

Empirical Likelihood Methods for Some Incomplete Data Problems

by

Menglu Che

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Statistics

Waterloo, Ontario, Canada, 2020

© Menglu Che 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jing Qin
Mathematical Statistician, Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases

Supervisor(s): Peisong Han
Assistant Professor, Department of Biostatistics,
University of Michigan
Jerald F. Lawless
Professor Emeritus

Internal Member: Changbao Wu
Professor
Pengfei Li
Professor

Internal-External Member: Pierre Chaussé
Associate Professor, Department of Economics

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Incomplete data often brings difficulty to estimations and inferences. A complete case (CC) analysis, in most cases, leads to biased estimates, or it may not have the desired estimation efficiency. In this thesis, we develop statistical methods addressing the estimation of regression parameters with missing covariates. We are interested in improving the estimation efficiency by incorporating the information from the partially observed cases.

Chapter 1 is an introduction to incomplete data problems and some existing estimation frameworks. We present the major tool we utilize to improve the estimation efficiency, i.e., empirical likelihood for general estimating functions. A brief introduction to the problems we solve in the subsequent chapters is also provided.

Chapter 2 considers a regression problem with covariates missing not at random, where the missingness depends on the missing covariate values. For this type of missingness, CC analysis leads to consistent estimation when the missingness is independent of the response given all covariates, but it may not have the desired level of efficiency. We propose a general empirical likelihood framework to improve the estimation efficiency upon CC analysis. We expand on methods in Bartlett, Carpenter, Tilling & Vansteelandt (2014) and Xie & Zhang (2017). Instead of improving the efficiency by modelling the missingness probability conditional on the response and fully observed covariates, our method allows the possibility of modelling other data distribution-related quantities. We also give guidelines on what quantities to model and demonstrate that our proposal has the potential to yield smaller biases than existing methods when the missingness probability model is incorrect. Simulation studies are presented, as well as an application to data collected from the US National Health and Nutrition Examination Survey.

Chapters 3 and 4 concern another type of incomplete data, namely the two-phase, response-dependent or outcome-dependent sample. This type of sampling is often used

in regression settings that involve expensive covariate measurements. Conditional maximum likelihood (CML) is an attractive approach in many cases as it avoids modelling the covariate distribution, unlike full maximum likelihood. Moreover, it handles zero selection probabilities of the Phase 2 sampling. In Chapter 3, we consider general regression models with either a discrete or continuous response. We show that the estimator of covariate effects proposed by Scott & Wild (2011) has the same asymptotic efficiency as two empirical likelihood estimators, and that these estimators dominate the CML estimator. Chapter 4 proposes a more general empirical likelihood method within the CML framework to incorporate the information in the Phase 1 sample and improve estimation efficiency. The proposed method exploits a model which only involves the fully observed variates. It maintains the ability to handle zero selection probability and avoids modelling the covariate distribution. The proposed methods exhibit improvement upon CML as well as the estimator by Scott & Wild (2011) considered in Chapter 3. In these two chapters, we compare the efficiencies of various estimators in simulation studies and illustrate the methodologies in a two-phase genetics study.

Chapter 5 presents some additional discussion and some topics for future research. We summarize the key points in our framework utilizing auxiliary information to improve estimation efficiency. Some additional remarks are given on the issues of numerical implementation, model diagnosis, and model compatibility. Finally, we discuss some topics for future research that are related to the methods considered in the thesis.

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisors, Drs. Peisong Han and Jerry Lawless, without whom the completion of my thesis would not be possible. I greatly appreciate their guidance, support, and inspiration through my doctoral research and writing the thesis. They would continue to be my lifelong role models of knowledge, professionalism and attitude towards academic research.

I would like to thank the examining committee, Drs. Jing Qin, Changbao Wu, Pengfei Li and Pierre Chaussé for reading my thesis and provide valuable comments to help improve the quality of it.

I also owe a debt of thanks to the Department of Statistics and Actuarial Science at the University of Waterloo who has provided funding and administrative services during my study. My special thanks go to my colleagues and friends at the Department, who supported me in various ways to continue my doctoral study. Last and most importantly, I would thank my husband, Jian's unconditional support during the years. To him I dedicate the thesis.

Dedication

This thesis is dedicated to Jian. For all his love, understanding and encouragement.

Table of Contents

List of Figures	xii
List of Tables	xiii
1 Introduction	1
1.1 Incomplete Data	1
1.2 Handling Different Types of Missing Data: an Analytical Introduction	2
1.3 Two-Phase, Outcome-Dependent Samples	9
1.4 The Empirical Likelihood Method for General Estimating Equations	12
1.5 Empirical Likelihoods for Incomplete Data	15
1.6 Contribution of the Thesis	17
2 Regression with MNAR Covariates	19
2.1 Introduction	19
2.2 Setup and Literature Review	22
2.3 A General Estimation Framework	25
2.4 Choices of Quantities to Model	28

2.5	Simulation Studies	31
2.5.1	Study 1	33
2.5.2	Study 2	36
2.6	Data Application	40
2.7	Discussion	43
3	Empirical and Conditional Likelihoods for Two-Phase, Outcome-Dependent Samples	45
3.1	Introduction	45
3.2	Estimators and Asymptotic Variances	48
3.2.1	The Conditional Maximum Likelihood (CML) Estimator	48
3.2.2	The Scott-Wild (SW) Estimator	50
3.2.3	The Empirical Likelihood Estimator	52
3.3	Simulation Studies	56
3.3.1	Simulation Study 1	56
3.3.2	Simulation Study 2	58
3.4	Illustration	60
3.5	Concluding Remarks	63
4	Improving the Efficiency for Estimation with Two-Phase, Outcome-Dependent Samples	64
4.1	Introduction	64
4.2	Theory and Methods	67

4.2.1	Data and Model Setup	67
4.2.2	Positive Selection Probability	68
4.2.3	When Zero Selection Probability for Certain Individuals is Present . .	74
4.3	Simulation Studies	78
4.3.1	Binary Outcome in a Logistic Regression Model with Expensive Co- variate	78
4.3.2	Continuous Outcome in a Linear Regression Model with Expensive Covariate	81
4.4	Illustration on a Genetics Study	84
4.5	Discussion	86
5	Discussions and Future Work	89
5.1	Discussion	89
5.1.1	Regression with Missing Covariates	89
5.1.2	Empirical Likelihood Frameworks for Exploiting Auxiliary Information	90
5.1.3	Numerical Implementation of Empirical Likelihood for General Esti- mating Equations	92
5.1.4	Model Checking and Model Compatibility	93
5.2	Future Work	94
	References	98
	APPENDICES	111

A Detailed Proofs of Theorems in Chapter 2	112
A.1 Proof of Theorem 1	112
A.2 Lemma 2 and Proof	115
B Additional Derivations and Numerical Results of Chapter 3	118
B.1 Derivations for relationships between the matrices \mathcal{C} and \mathcal{J}	118
B.2 Calculation of the asymptotic variance of the EL estimator	124
B.3 The rank of CL estimating equations for Studies 1 and 3	125
B.4 Additional simulation studies	128
B.4.1 Simulation study 3	128
B.4.2 Simulation study 4	130
C Additional Derivations and Numerical Results of Chapter 4	132
C.1 Proofs of (4.6)	132
C.2 Equivalence of the EL0-1 and EL1 estimator	134
C.3 Equivalence of the EL0-2 and EL2 estimator	134
C.4 Additional simulation study 3: a logistic regression model with surrogate covariate	142
C.5 Additional simulation study 4: normal linear regressions with varying covari- ate correlation	144
Glossary	146

List of Figures

2.1	P-P plot for the three models for $f(Z Y, X, W, R = 1)$ with $n = 400$	38
-----	--	---------	----

List of Tables

2.1	Simulation results for Study 1.	35
2.2	Simulation results for Study 2.	39
2.3	Analysis results for the NHANES data	42
3.1	Simulation results for Study 1.	59
3.2	Simulation results for Study 2.	61
3.3	Regression coefficient estimates for the GAW 17 two-phase data	62
4.1	Binary outcome in a logistic regression model, $n = 2000$	81
4.2	Continuous outcome in a linear regression model with positive selection probabilities, $n = 2000$	84
4.3	Continuous outcome in a linear regression model with positive selection probabilities, $n = 300$	85
4.4	Continuous outcome in a linear regression model with zero selection probabilities, $n = 2000$	85
4.5	Continuous outcome in a linear regression model with zero selection probabilities, $n = 300$	86
4.6	Regression coefficient estimates for the GAW 17 two-phase data	87

B.1	Simulation results for Study 3.	130
B.2	Simulation results for Study 4.	131
C.1	Relative efficiencies for logistic regression models with surrogate covariates .	143
C.2	Relative efficiencies for a linear regression model, with $\beta_{c0} = 0$, $\beta_{z0} = 1$, $\beta_{x0} = 1$, $\sigma_0 = 2$	145
C.3	Relative efficiencies for a linear regression model, with $\beta_{c0} = 0$, $\beta_{z0} = 0.5$, $\beta_{x0} = 0.5$, $\sigma_0 = 2$	145

Chapter 1

Introduction

Incomplete data is ubiquitous in research areas such as survey sampling, social and medical sciences. Most of the time, it brings difficulty to the estimation and inference procedures of the target study. This thesis focuses on improving the estimation efficiency of general regression models with incomplete data through semiparametric methods, especially empirical likelihood, for estimating equations. In this chapter, we give a general introduction to regression problems with incomplete covariates, empirical likelihood, and how they are linked to the specific problems in the following chapters.

1.1 Incomplete Data

Incomplete data means the absence of measurements in data collection. These absences may be intentional or unintentional. Intentional incomplete data often comes from the deliberate design of sampling plans. For example, due to the restriction of sampling costs, researchers often cannot afford to measure an expensive variable for all subjects in a representative sample. Other incomplete data may come from factors that are out of researchers'

control, such as nonresponse to questionnaires, being unable to measure some variates, and data loss.

The aforementioned incomplete data share a common characteristic that there is a partial loss of information. In the past few decades, tremendous efforts have been made to provide analytical strategies to properly account for incomplete data (Little & Rubin 2014). The type of incompleteness that has been studied most thoroughly is missing data, where data value is either perfectly known or entirely unknown. Rubin (1976) presents a general model of missing data treating the missingness indicator as a random variable and assigning it a distribution. Statistical analysis must take account of the mechanisms that give rise to the missingness. Improper analysis of the incomplete data may lead to bias or loss of statistical power. For example, the complete case (CC) analysis which simply ignores the individuals with missing observations, leads to biased estimates most of the time.

1.2 Handling Different Types of Missing Data: an Analytical Introduction

This thesis is devoted to the analysis of regression problems with a scalar response and part of the covariates subject to missingness. For a data set of $i = 1, \dots, n$ subjects, we denote the complete data as $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, where i indexes the subject. Let Y_i stand for the response, and $(\mathbf{X}_i, \mathbf{Z}_i)$ stand for the covariates, where \mathbf{Z}_i may be missing for part of the data set. The missingness indicator is defined as

$$R_i = \begin{cases} 1, & \text{if } \mathbf{Z}_i \text{ is observed,} \\ 0, & \text{if } \mathbf{Z}_i \text{ is missing.} \end{cases}$$

The observed data is thus $(Y_i, \mathbf{X}_i, R_i \mathbf{Z}_i, R_i)$. As defined by Little & Rubin (2014), the mechanisms of missing data can be categorized into three types: missing completely at random ([MCAR](#)), missing at random ([MAR](#)), and missing not at random ([MNAR](#)). The first type, MCAR, accounts for missing data that does not depend on any observed or unobserved quantities. That is,

$$f(R|Y, \mathbf{X}, \mathbf{Z}; \phi) = f(R; \phi),$$

where ϕ is a parameter vector which can be known or unknown. Complete-data-only inference is valid under this missingness mechanism, though it may not be most efficient.

MAR data describes the mechanism where the missingness probability depends on the observed quantities only and may vary across different individuals. That is,

$$f(R|Y, \mathbf{X}, \mathbf{Z}; \phi) = f(R|Y, \mathbf{X}, R\mathbf{Z}; \phi).$$

MCAR is a special case of MAR, but in MAR cases, complete case (CC) analysis is not valid in general and may produce biased results. In some cases, a scenario that falls between MAR and MCAR is the covariate-dependent-missingness ([CDM](#)) mechanism, that is, when the missingness probability only depends on the fully observed variates but not the partially observed variates, i.e.,

$$f(R|Y, \mathbf{X}, \mathbf{Z}; \phi) = f(R|Y, \mathbf{X}; \phi).$$

We may use CDM interchangeably with MAR in later sections and chapters.

The concept of ignorability is also commonly used in missing data literature. If we are interested in a parameter θ that indexes the complete data, under the MAR assumption we

may write

$$f(Y, \mathbf{X}, \mathbf{Z}, R; \boldsymbol{\theta}, \phi) = f(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})f(R|Y, \mathbf{X}, \mathbf{Z}; \phi) = f(\mathbf{Y}; \boldsymbol{\theta})f(R|Y, \mathbf{X}, R\mathbf{Z}; \phi),$$

so the likelihood of $\boldsymbol{\theta}$ and ϕ is

$$L_{\text{full}}(\boldsymbol{\theta}, \phi) \propto f(Y, \mathbf{X}, R\mathbf{Z}, R; \boldsymbol{\theta}, \phi) = f(Y, \mathbf{X}, R\mathbf{Z}; \boldsymbol{\theta})f(R|Y, \mathbf{X}, R\mathbf{Z}; \phi). \quad (1.1)$$

When the parameters $\boldsymbol{\theta}$ and ϕ are distinct, the full likelihood can be factorized into two parts that have separate parameters, so the missingness mechanism can be ignored for inference about $\boldsymbol{\theta}$, and we say this type of missingness is *ignorable* for that purpose. In ignorable missing data, likelihood-based analysis on the observed data provides valid inference (Molenberghs & Kenward 2007).

When neither the MAR nor MCAR assumption holds, the data is said to be MNAR or non-ignorably missing. The missingness probability may vary among individuals and depend on the underlying missing variates. In practice, MNAR is often the most realistic assumption to impose, for example, when there are missing values for sensitive variates related to income or drug use. Nevertheless, MNAR problems are far more challenging to deal with than MAR problems. They usually require more complicated model assumptions and additional information to identify model parameters.

Various methods have been developed over the past few decades for statistical analysis with missing data. For all the methods other than complete case (CC) analysis, Schafer & Graham (2002) summarized these methods into two main approaches: maximum likelihood (ML) estimation based on all observed data, and Bayesian multiple imputation (MI). There is a third approach via using estimating equations, which does not necessarily use a full likelihood model. We shall discuss this in more details later. MI generates multiple sets of

imputed values for the missing variates, and yields complete data sets as well as corresponding estimates. Then these estimates may be combined to produce a final estimate. Examples of this approach include Rubin (1976), Rubin (1996), Raghunathan, Reiter & Rubin (2003) and Sterne, White, Carlin, Spratt, Royston, Kenward, Wood & Carpenter (2009). For ML methods there are two ways to specify the likelihood. One is the so-called *selection model*:

$$f(R, Y | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}, \phi) = f(Y | \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) f(R | Y, \mathbf{X}, \mathbf{Z}; \phi). \quad (1.2)$$

The alternative model of the data is

$$f(R, Y | \mathbf{X}, \mathbf{Z}; \boldsymbol{\gamma}, \boldsymbol{\psi}) = f(Y | \mathbf{X}, \mathbf{Z}, R; \boldsymbol{\gamma}) f(R | \mathbf{X}, \mathbf{Z}; \boldsymbol{\psi}), \quad (1.3)$$

which is referred to as the *pattern mixture model*.

The main focus of this thesis is on the usage of estimating equations to model the marginal or population averaged relationship between a response variate and predictors (Zeger, Liang & Albert 1988). For a sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \dots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ indexed by a parameter $\boldsymbol{\theta}$, an estimating equation is in the form

$$\sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) = \mathbf{0} \quad (1.4)$$

where $\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$ is any function of $Y, \mathbf{X}, \mathbf{Z}$ indexed by a parameter $\boldsymbol{\theta}$. When

$$E_{(Y, \mathbf{X}, \mathbf{Z})} \{ \mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_0) \} = \mathbf{0}$$

for some fixed $\boldsymbol{\theta}_0$, the estimating equation is said to be *unbiased*. For example, for a multiple linear regression model, with $E(Y | \mathbf{X}, \mathbf{Z}) = \boldsymbol{\theta}^T(\mathbf{X}, \mathbf{Z})$, the least square estimator $\hat{\boldsymbol{\theta}}_{\text{LS}}$

corresponds to the estimating equation

$$\sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) = \sum_{i=1}^n (\mathbf{X}_i, \mathbf{Z}_i) \{Y_i - \boldsymbol{\theta}^T(\mathbf{X}_i, \mathbf{Z}_i)\} = \mathbf{0}.$$

Estimating equations play a vital role in semiparametric methods (Tsiatis 2007) yielding the so-called Z-estimator. For parametric models, an ML estimator is also equivalent to solving an estimating equation with the score function equal to zero. We now describe some popular estimating-equation-based estimators for a regression model concerning $Y|\mathbf{X}, \mathbf{Z}$.

With a sample $\{(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \dots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)\}$, the CC analysis corresponds to solving the CC estimating equation

$$\sum_{i=1}^n R_i \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) = \mathbf{0}, \quad (1.5)$$

and $\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$ can be any estimating function. It is easy to see that for MCAR data, R is independent of Y , \mathbf{X} and \mathbf{Z} , hence the CC analysis provides valid inference, or consistent estimate of the parameter $\boldsymbol{\theta}$. CC is also valid when $R \perp Y|\mathbf{X}, \mathbf{Z}$. In both cases, we have

$$E\{R\mathbf{U}(YR, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})|\mathbf{X}, \mathbf{Z}, R=1\} = E\{\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})|\mathbf{X}, \mathbf{Z}\}. \quad (1.6)$$

However, generally speaking, solving (1.5) yields biased estimate of $\boldsymbol{\theta}$.

Based on only the fully observed (complete) data, reweighting is a popular method to handle MAR data. Initiated by the Horvitz-Thompson estimator in survey sampling, Rosenbaum & Rubin (1983) first officially termed the method as inverse probability weighting (IPW). IPW reweights each case by the estimated, or known missingness probability. The

estimating equation can be written as $\mathbf{S}_{\text{IPW}}(\boldsymbol{\theta}) = \mathbf{0}$ where

$$\mathbf{S}_{\text{IPW}}(\boldsymbol{\theta}) = \sum_{i=1}^n R_i w_i \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) \quad (1.7)$$

with $1/w_i = \pi(Y_i, \mathbf{X}_i) = P(R_i = 1|Y_i, \mathbf{X}_i) = P(R_i = 1|Y_i, \mathbf{X}_i, \mathbf{Z}_i)$. By this reweighting approach, IPW corrects the bias of CC. When w_i 's are unknown, and the data is MAR, w_i 's can be estimated by indexing the selection model by a nuisance parameter $\boldsymbol{\alpha}$, i.e., $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$. Provided the $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ model is correctly specified, IPW yields consistent estimates.

An alternative approach is to work with $f(Y|\mathbf{X}, \mathbf{Z}, R = 1)$ instead of $f(Y|\mathbf{X}, \mathbf{Z})$, which leads to solving an estimating equation $\mathbf{S}_{\text{CML}}(\boldsymbol{\theta}) = \mathbf{0}$ where $\mathbf{S}_{\text{CML}}(\boldsymbol{\theta})$ stands for the conditional maximum likelihood (CML) score function:

$$\mathbf{S}_{\text{CML}}(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{\partial \log f_c(Y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \log \left\{ \frac{f(Y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) \pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})}{\int f(y|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) \pi(y, \mathbf{X}_i; \boldsymbol{\alpha}) dy} \right\}.$$

Though IPW corrects the bias in ideal cases, it fails when the model for the missingness, $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ is misspecified. The IPW estimator uses only the complete cases and may lose efficiency in this sense. Robins, Rotnitzky & Zhao (1994) proposed an improved version, the augmented inverse probability weighting (AIPW) approach. The AIPW estimator augments the estimating function of IPW by an auxiliary estimating function, which utilizes the information in all partially observed cases. See Seaman & Vansteelandt (2018) for an comprehensive introduction. For our missing covariate problem, the estimating equation of AIPW is of the form

$$\mathbf{S}_{\text{AIPW}}(\boldsymbol{\theta}) = \sum_{i=1}^n \left[\frac{R_i}{\pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})} \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta}) + \left\{ 1 - \frac{R_i}{\pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})} \right\} \boldsymbol{\phi}(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) \right] = \mathbf{0} \quad (1.8)$$

where $\boldsymbol{\phi}(Y, \mathbf{X}; \boldsymbol{\theta})$ is a function of the always observed variables. An optimal form of

$\phi(Y, \mathbf{X}; \boldsymbol{\theta})$ is the estimate of the expectation of $\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$ given observed data $(Y, \mathbf{X}, R\mathbf{Z})$, which can be viewed as an imputed value for the original estimating function $\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})$. It is only required that the missingness only depends on the fully observed variates Y, \mathbf{X} . Thus, AIPW is a hybrid of IPW and imputation. AIPW is known to possess the favourable double robustness (DR) property: provided either the missingness model or the imputation model is correctly specified, i.e., $E\{R - \pi(Y, \mathbf{X}, \boldsymbol{\alpha})\} = 0$ or $E\{\phi(Y, \mathbf{X}; \boldsymbol{\theta})\} = \mathbf{0}$, the AIPW estimator is consistent. When both models are correct, the AIPW estimator achieves the semiparametric efficiency bound. We refer to Kang, Schafer et al. (2007) as a comprehensive reference discussing the DR property.

In popular methods for missing data, the primary estimating function which identifies the parameter of interest may not fully utilize all the information contained in the observed data. Often, estimation efficiency may be improved by exploiting the partially observed cases or certain nuisance parameters. For example, the AIPW estimator utilizes the partially observed cases in the augmentation part to improve upon IPW. In the case \mathbf{Z} is MNAR with $R \perp Y | \mathbf{X}, \mathbf{Z}$, Bartlett et al. (2014) improves the efficiency of CC analysis by adding another estimating function component of Y, \mathbf{X} and R . Multiple estimating functions, for both the parameter of interest and nuisance parameters, naturally arises from this auxiliary information, and often times, more estimating functions than free parameters can be defined. We refer the readers to Qin (2017) as a comprehensive review of such *over-identified* parameter problems. For solving these types of problems, empirical likelihood serves as a powerful tool with favourable properties and thus is a popular approach to cope with missing data. We will later use Section 1.5 to describe the empirical likelihood method.

1.3 Two-Phase, Outcome-Dependent Samples

One type of incomplete data extensively studied in this thesis is the two-phase, outcome-dependent samples (ODS). Two-phase samples can be dated back to Neyman (1938). As the name suggests, the sampling occurs in two phases. In Phase 1, one obtains the values of certain variables on a simple random sample or entire cohort, and stratifies the sample into strata. Then in Phase 2, a random sub-sample of individuals is taken to have other detailed information measured. The selection of the Phase 2 sample often depends on the Phase 1 sample. This approach is particularly desirable when the occurrence of some variables is rare, or part of the covariates are expensive to measure. It is found that a Phase 2 sample overrepresenting the rare or extreme outcomes may drastically improve the estimation efficiency (Breslow 1996; Huang & Lin 2007; Chen & Li 2011). For example, for binary outcome representing a rare disease incidence, epidemiologists often sample most of the disease cases with $Y = 1$ and only a small fraction of the controls with $Y = 0$ (Breslow & Holubkov 1997a). For a continuous outcome, extreme-tail samples are often used, which only samples the individuals with outcome or exposure values on the two end tails (Huang & Lin 2007; Lin, Zeng & Tang 2013). This type of two-phase sample is then said to be “outcome-dependent” as it depends on the outcome values.

In particular, we study a regression model with outcome Y_i and covariate vector \mathbf{X}_i that can be routinely measured for a representative sample $i = 1, \dots, n$ in Phase 1, and an expensive covariate vector, denoted as \mathbf{Z}_i here, is measured for a Phase 2 sub-sample $i = 1, \dots, m$. With a parametric model of interest, $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$, indexed by parameter $\boldsymbol{\beta}$, one can write the joint density of (Y_i, \mathbf{Z}_i) given \mathbf{X}_i as $f(Y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})g(\mathbf{Z}_i|\mathbf{X}_i)$. To make the estimation of $\boldsymbol{\beta}$ more efficient, the Phase 2 sample is based on the response values observed in Phase 1. Well-known examples of ODS include case-control and case-cohort studies used with rare outcomes and two-phase studies stratified on a continuous outcome

and/or inexpensive covariate measured in Phase 1. Compared to simple random sampling, two-phase samples may be much more cost-effective and are widely used in large-scale studies. See, for example, Holcroft, Rotnitzky & Robins (1997), Lawless, Kalbfleisch & Wild (1999) and Breslow, Lumley, Ballantyne, Chambless & Kulich (2009). Since the Phase 2 selection by design only depends on the response and covariate variables whose values are collected in Phase 1, using an indicator variable R_i to denote if the i -th individual is included in Phase 2, a selection model can be written as $\pi_i(\boldsymbol{\alpha}) = \pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha}) = P(R_i = 1 | Y_i, \mathbf{X}_i; \boldsymbol{\alpha})$, indexed by a nuisance parameter $\boldsymbol{\alpha}$. The correct model of $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ and the true value $\boldsymbol{\alpha}_0$ of $\boldsymbol{\alpha}$ is then known by design. Therefore, the observed samples are $(Y_i, \mathbf{X}_i, R_i \mathbf{Z}_i, R_i)_{i=1}^n$ which are iid and satisfy $R \perp \mathbf{Z} | Y, \mathbf{X}$. The ODS problem is then straightforwardly framed into a missing-at-random (MAR) problem in the context of missing data, where we treat the \mathbf{Z}_i values with $R_i = 0$ as missing.

With the model $f(Y | \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$, various methods for such two-phase studies have been discussed in a large body of works, see, for example, (Breslow & Holubkov 1997b) and Lawless et al. (1999). Generally, candidate methods for analyzing two-phase, ODS data falls into four categories. The first and most efficient one is full maximum likelihood (ML), by maximizing the joint likelihood

$$L(\boldsymbol{\beta}) = \prod_{i \in V} f(Y_i | \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) \prod_{j \in \bar{V}} \int f(Y_j | \mathbf{X}_j, \mathbf{z}; \boldsymbol{\beta}) dG(\mathbf{z} | \mathbf{X}_j), \quad (1.9)$$

where V and \bar{V} represents the indices of the fully observed and partially observed cases, respectively, and $G(\mathbf{z} | \mathbf{X}_j)$ is the cumulative distribution function corresponding to the density $g(\mathbf{z} | \mathbf{X}_j)$. The covariate distribution $G(\cdot | \mathbf{X})$ can be modelled either parametrically or nonparametrically. The second, often referred to as pseudolikelihood methods, uses estimating equations related to maximum likelihood, usually a score-like function from a certain log-pseudolikelihood. For example, for one-dimensional Z and categorical X , one can use

empirical estimate $G_N(z|x, R = 1) = G_N(z|x, R = 1) := \sum_i I(Z_i \leq z, X_i = x, R_i = 1) / \sum_i I(X_i = x, R_i = 1)$ to estimate $G(z|X = x, R = 1)$ and write out the estimated $d\hat{G}(z|X) = dG_N(Z_i \leq z|X, R = 1)P(R = 1|X)/P(R = 1|X, Z)$ in 1.9 (Chatterjee, Chen & Breslow 2003). The third is the inverse probability weighting (IPW) method based on the Horvitz-Thompson (HT) estimator, by assuming positive selection probability in phase 2 for any individual and solving $\sum_i R_i \mathbf{s}_\beta(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta) / \pi(Y_i, \mathbf{X}_i) = \mathbf{0}$ where \mathbf{s}_β is the score function $\partial \log\{f(Y|\mathbf{X}, \mathbf{Z}; \beta)\} / \partial \beta$. The last one is the “complete-data likelihood” or conditional likelihood. It relies on the weighted distribution of the phase 2 sample, based on both the response model $f(Y|\mathbf{X}, \mathbf{Z}; \beta)$ and the selection model $\pi(Y, \mathbf{X}) = P(R = 1|Y, \mathbf{X})$, namely

$$f_c(Y|\mathbf{X}, \mathbf{Z}, R = 1; \beta) = \frac{f(Y|\mathbf{X}, \mathbf{Z}; \beta)\pi(Y, \mathbf{X})}{\int f(y|\mathbf{X}, \mathbf{Z}; \beta)\pi(y, \mathbf{X})dy}, \quad (1.10)$$

and maximizing the corresponding likelihood. We follow Breslow, Zhao, Fears & Brown (1988) and Scott & Wild (2011) among others and name the estimator as conditional maximum likelihood (CML) estimator.

The estimators mentioned above are not disjoint with each other. For example, in the special case of case-control studies with a binary outcome, CML is ML for the odds-ratio parameter; when the IPW estimator uses a score function of all phase 2 data, it is also viewed as a weighted pseudo-likelihood estimator. All the estimators are consistent when the involving parametric parts are correctly specified. A variety of classic estimators are based only on the phase 2 sample, i.e., the complete cases, including CML and IPW. While ML might be difficult or impossible to use in certain scenarios, utilizing the information carried by the Phase 1 sample is a popular topic, see Scott & Wild (2011), Rivera-Rodriguez, Haneuse, Wang & Spiegelman (2020) among others.

1.4 The Empirical Likelihood Method for General Estimating Equations

Empirical likelihood was first proposed by Owen (1988) to estimate the parameters of distributions. It is a semiparametric tool for estimation of parameters, but has attractive characteristics similar to parametric likelihood. More specifically, empirical likelihood gives confidence regions just as maximum (parametric) likelihood, and the likelihood ratio statistic also follows a Chi-square distribution asymptotically, making it possible for hypothesis testing. Owen (1990) discussed empirical likelihood as an alternative method for bootstrap. Hall & La Scala (1990) discussed some advantages of the empirical likelihood approach over bootstrap. For example, in the estimation of a mean, we do not need to estimate a scale or skewness parameter to get a confidence region. The empirical likelihood confidence regions are determined totally by the sample distribution, thus reflects any special characteristics of this distribution. The regions are range preserving and transform preserving just like MLE. Moreover, the regions are Bartlett correctable (DiCiccio, Hall & Romano 1991), with a simple correction reducing the coverage error from order n^{-1} to n^{-2} . We refer the readers to Owen (2001) for a comprehensive review and discussion of the empirical likelihood method.

The work of Qin & Lawless (1994) made empirical likelihood more powerful for more general semiparametric problems, by combining empirical likelihood and general estimating equations. It provides an alternative to the generalized method of moments (GMM) for solving estimation problems without fully specifying the likelihood (Hansen 1982). Both GMM and empirical likelihood are suitable especially for over-identified estimating equations. Later, Newey & Smith (2004) showed that asymptotically, GMM and empirical likelihood behave equivalently to the first order, but empirical likelihood has better higher-order properties.

For a random variable \mathbf{Y} and a p -dimensional parameter of interest, $\boldsymbol{\theta}$, the empirical likelihood method can be applied to a set of r ($r \geq p$) functionally independent general estimating functions $g_j(\mathbf{Y}; \boldsymbol{\theta})$, satisfying

$$E_{\mathbf{Y}}\{\mathbf{g}(\mathbf{Y}; \boldsymbol{\theta})\} = E\{g_1(\mathbf{Y}; \boldsymbol{\theta}), g_2(\mathbf{Y}; \boldsymbol{\theta}), \dots, g_r(\mathbf{Y}; \boldsymbol{\theta})\} = \mathbf{0}.$$

With a random sample $\mathbf{y}_1, \dots, \mathbf{y}_n$ for the variate of interest, \mathbf{Y} , we solve the following constrained optimization problem:

$$\begin{aligned} \max_{p_i, \boldsymbol{\theta}} \prod_{i=1}^n p_i &= \prod_{i=1}^n dF(\mathbf{y}_i), \text{ subject to} \\ p_i &> 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta}) = \mathbf{0}. \end{aligned}$$

This constrained optimization is solved through Lagrange multipliers. The Lagrange function is

$$\mathcal{L} = \sum_{i=1}^n \log(p_i) + \mu(1 - \sum_{i=1}^n p_i) + n\boldsymbol{\lambda}^T \left\{ \sum_{i=1}^n p_i \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta}) \right\},$$

where $\boldsymbol{\lambda}$ and μ are Lagrange multipliers. Specifically, $\boldsymbol{\lambda}$ has dimension r as $\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})$ is an r -dimensional vector. Taking derivative with respect to p_i , we get

$$\frac{\partial \mathcal{L}}{\partial p_i} = \frac{1}{p_i} - \mu + n\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta}) = 0,$$

and

$$\sum_{i=1}^n p_i \frac{\partial \mathcal{L}}{\partial p_i} = n - \mu = 0,$$

so $\hat{\mu} = n$ and $\hat{p}_i = \{n - n\boldsymbol{\lambda}^T \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta})\}^{-1}$, and we can write the original optimization as maximizing the profile log empirical likelihood,

$$l(\boldsymbol{\theta}) = - \sum_{i=1}^n \log[1 + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta})] \quad (1.11)$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\lambda}$ is determined by

$$\sum_{i=1}^n \frac{\mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{y}_i; \boldsymbol{\theta})} = 0.$$

The resulting estimate, $\hat{\boldsymbol{\theta}}_{\text{EL}}$ is called the maximum empirical likelihood estimate (MELE). When all the estimating functions are unbiased at a unique $\boldsymbol{\theta}_0$, under some regularity conditions, the estimate $\hat{\boldsymbol{\theta}}_{\text{EL}}$ is consistent and asymptotically normal. It is fully efficient in the sense that it has the same asymptotic variance as the optimal estimator obtained from the class of $p \times 1$ estimating equations that are linear combinations of $g_1(\mathbf{Y}; \boldsymbol{\theta}), g_2(\mathbf{Y}; \boldsymbol{\theta}), \dots, g_r(\mathbf{Y}; \boldsymbol{\theta})$. Specifically,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{EL}} - \boldsymbol{\theta}_0) \Rightarrow N(0, V)$$

with

$$V = \left[E \left\{ \frac{\partial \mathbf{g}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \right\} E \{ \mathbf{g}(\boldsymbol{\theta}_0) \mathbf{g}(\boldsymbol{\theta}_0)^T \}^{-1} E \left\{ \frac{\partial \mathbf{g}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\} \right]^{-1}. \quad (1.12)$$

By Qin & Lawless (1994), if part of the estimating equations are dropped, the variance estimate of the estimator will not decrease.

1.5 Empirical Likelihoods for Incomplete Data

In the estimation and inference for incomplete data, some authors use imputation for the estimation, while using empirical likelihood to construct confidence intervals (Wang & Rao 2002; Liang, Wang & Carroll 2007); some authors use empirical likelihood to directly get the estimate of the parameter as well as the estimated variance. Qin, Zhang & Leung (2009) provided a collection of examples of missing data problems under the assumption of MAR, and showed how to utilize empirical likelihood in these types of problems, including estimation and inference with missing covariates in regression model and with surrogate responses. In these settings, the regression model is of interest and the missingness mechanism is only auxiliary. Qin et al. (2009) used the components of the AIPW estimating functions and proposed several options to combine them into over-identified estimating functions. Specifically, for a regression problem $Y = \mu(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}) + \epsilon$ with error ϵ , where \mathbf{Z} is missing at random, they write the estimating equation in (1.8) as $\mathbf{g}_1 - \mathbf{g}_2 = \mathbf{0}$, where

$$\mathbf{g}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n \left[\frac{R_i}{\pi(Y, \mathbf{X}; \boldsymbol{\alpha})} \mathbf{U}(Y, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) \right],$$

$$\mathbf{g}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n \left[\left\{ 1 - \frac{R_i}{\pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})} \right\} \boldsymbol{\phi}(Y_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) \right].$$

$\boldsymbol{\gamma}$ is a nuisance parameter to index the auxiliary function $\boldsymbol{\phi}(Y_i, \mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})$. In most cases, the nuisance parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ can be estimated independently without estimating $\boldsymbol{\beta}$. For example, $\boldsymbol{\alpha}$ can be estimated by maximizing the likelihood

$$\prod_{i=1}^n \pi(Y, \mathbf{X}; \boldsymbol{\alpha})^{R_i} \{1 - \pi(Y, \mathbf{X}; \boldsymbol{\alpha})\}^{1-R_i}.$$

Therefore, we can write an estimating equation of only the nuisance parameters, which we denote as $\mathbf{g}_3(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathbf{0}$. Qin et al. (2009) proposed to estimate the nuisance parameter first, and fix it during the following empirical likelihood estimation of the parameter of interest, i.e., using $\mathbf{g}_j(\boldsymbol{\beta}) = \mathbf{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}), j = 1, 2$; or $\mathbf{g}_j(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}), j = 1, 2$. The resulting estimator is called the maximum pseudo-empirical likelihood estimator (PMELE). Although asymptotically PMELE and MELE are equivalent, for finite samples, they have slightly different variance estimates.

A further improved method rooting from the idea of AIPW is the multiply robust (MR) estimator (Han & Wang 2013; Han 2014; Chan 2013; Han 2016). For regression models where the outcome Y is subject to missingness, the MR estimator uses the idea of empirical likelihood and allows multiple choices of both the missingness model and the data distribution model. Given the class of missingness models and the class of data distribution models, when either class of models has a correct model included, the MR estimator is consistent. Thus it provides better robustness than the DR estimator.

Compared to ignorably missing data problems, nonignorably missing data problems face a much more complicated likelihood structure. The missingness probability and the underlying response model of interest may not be factored out, thus must be handled simultaneously, except for some very special cases (e.g., Bartlett et al. 2014). There are a few works on likelihood-based inference for MNAR data. For the parametric likelihood case, Miao, Ding & Geng (2016) studied identifiability for normal mixture models; for semiparametric problems, Tang, Zhao & Zhu (2014) studied an empirical likelihood assisted imputation method for estimating equations with nonignorably missing data. Chapter 22 of Qin (2017) is devoted to general problems with MNAR data.

1.6 Contribution of the Thesis

In this thesis, we focus on missing covariate problems in regression models but consider different model assumptions, including both the variables' distributions and the missingness mechanisms. There is a variety of literature available studying such problems, but generally, they can be categorized by the regression type, for example, linear regression vs logistic regression, or different missingness mechanisms, for example, MAR vs. MNAR.

In Chapter 2, we address a special case of MNAR covariates, where CC provides a valid estimate of the regression coefficients, but may not be efficient. Our task is to utilize the partially observed cases to improve the efficiency of CC. The approach is inspired by the work by Bartlett et al. (2014) and Xie & Zhang (2017). We find that their proposed method is restrictive in the additional model assumptions, and when they are violated the method is prone to bias. Instead, we propose a more general framework that allows more flexible model assumptions. We show that in general, simultaneous estimation of auxiliary models improves the efficiency of CC, and the methods in Bartlett et al. (2014) and Xie & Zhang (2017) turn into special cases of our framework.

Conditional maximum likelihood is an attractive approach for MAR covariates as it avoids the modelling of the covariate distribution. Scott & Wild (2011) finds the specific form of a semiparametric efficient estimator for a binary outcome and categorical covariates, and showed that it may be extended to more general forms of variables such as continuous covariates. We refer to this form of estimator as the SW estimator. The SW estimator's idea is that, when introducing the nuisance parameter α in the missingness model $\pi(Y, \mathbf{X}; \alpha)$, the conditional likelihood is also a function of α , thus it has a conditional score with respect to α , too. The score function serves as an augmentation part in the estimating equations, and may improve the estimation efficiency for the parameter of interest. We consider a general regression model for $f(Y|\mathbf{X}, \mathbf{Z}; \beta)$, and show an estimator of the same form as SW to be

asymptotically equivalent to two empirical likelihood estimators in terms of efficiency of the parameter of interest.

Chapter 4 presents a more general framework to use a very flexible auxiliary model to enhance the conditional maximum likelihood estimator for two-phase, outcome dependent samples. Practically, a regression model of Y given the always observed covariate \mathbf{X} is a natural choice. As both Y and \mathbf{X} are observed for the entire Phase 1 sample, the postulation and diagnosis of such a regression model would not be difficult. The framework enjoys all the advantages of conditional maximum likelihood. It allows a wide range of two-phase designs, which may involve zero selection probabilities for part of the Phase 1 sample. It also handles both continuous and discrete outcomes, as well as any types of covariates, continuous or discrete, one-dimensional or multi-dimensional.

We can unite the above estimators in a very general form of problems. For regression with incomplete covariates, when there is a CC, or complete-data-based estimator, we can employ the fully observed variables to enhance such an estimator using empirical likelihood. Discussions are given in Chapter 5, in which we also provide our outlook towards some promising extensions and future works relevant to the thesis.

Chapter 2

Improving Estimation Efficiency for Regression with MNAR Covariates

2.1 Introduction

Regression analysis is often complicated by the presence of missing data. Handling missing data inappropriately can lead to biased estimation and/or loss of efficiency. The most commonly used assumption about the missingness mechanism is missing-at-random (MAR), where the missingness depends only on the observed data but not on the missing data. There is a rich collection of effective methods dealing with MAR data, including multiple imputation (Rubin 1987), inverse probability weighting (Horvitz and Thompson 1952), augmented inverse probability weighting (Robins et al. 1994), and other likelihood-based methods (Little and Rubin 2002). However, in many settings, the assumption of MAR is too strong and the missingness does depend on the missing data even conditional on the observed data. Developing general methods dealing with such missing-not-at-random (MNAR) data is very challenging due to model identifiability. See, for example, Rotnitzky and Robins (1997),

Ibrahim et al. (1999), Wang et al. (2014), Miao and Tchetgen Tchetgen (2016), and Han (2018) for some relevant discussions.

In this chapter, we consider regression analysis with MNAR covariates where the missingness is assumed to be independent of the response given all covariates of interest. This is a practically important setting, especially when the covariates are measured at the beginning of the study but the response is measured at a later time point. In this case, it is natural and logical to assume that the missingness of covariates does not depend on the future response values once given all covariate values, but may depend on the covariates. For such a setting, a complete-case analysis based only on subjects with fully observed data leads to consistent estimation of the regression parameters. However, the complete-case analysis ignores the information in the partially observed subjects and thus may not have the desired level of efficiency, especially when the proportion of subjects with missing data is not small. How to effectively use the partially observed information to improve estimation efficiency over the complete-case analysis is of great interest.

By modelling the missingness given both the response and the subset of fully observed covariates, Bartlett et al. (2014) proposed the augmented complete-case (ACC) estimator. Note that the missingness model they assumed is not for the MNAR mechanism, which depends on the subset of missing covariates as well, but is rather for the distribution of the missingness indicator given all fully observed variables in the data set. With this model assumption, Bartlett et al. (2014) derived the optimal augmentation term that ensures an efficiency improvement over the complete-case analysis. Noting that the ACC estimating function is a simple sum of the complete-case analysis estimating function and an augmentation term, Xie and Zhang (2017) proposed to treat the two pieces as an over-identified estimating function and estimated the regression parameters based on the empirical likelihood method (Qin and Lawless 1994), which essentially finds the optimal linear combination of the two pieces instead of simply summing them up. Such an application of the empirical

likelihood method has also been considered for MAR data. See, for example, Qin et al. (2009).

Both Bartlett et al. (2014) and Xie and Zhang (2017) assumed a model for the missingness given all the fully observed variables to improve efficiency over the complete-case analysis. It may be possible to model other quantities to achieve the same goal. One straightforward example is, with the observed data, to model the distribution of the response given the subset of fully observed covariates. Note that this model is different from the regression model of primary interest that models the response given all covariates. It is natural to ask how to accommodate these different model assumptions into estimation and if they are also able to extract information from the partially observed subjects. In this chapter, we propose a general empirical likelihood-based framework for efficiency improvement that can accommodate different model assumptions. These assumptions yield extra estimating functions in addition to the ones used for the complete-case analysis. As a result, this general framework covers the methods in Bartlett et al. (2014) and Xie and Zhang (2017) when using their assumed model for the missingness. We also provide some guidelines on what quantities to model for a better efficiency improvement. Although a theoretical justification of these guidelines seems infeasible, they are formulated based on logical intuition and lead to good numerical performances in our simulation studies. As an illustration of the proposed method, we analyze data collected from the US National Health and Nutrition Examination Survey (NHANES).

The rest of this chapter is organized as follows. Section 2.2 gives the setup and a review of relevant methods. Section 2.3 covers the proposed general framework. Section 2.4 provides some guidelines on what quantities to model to have a better efficiency improvement. Sections 2.5 and 2.6 contain simulation studies and a data application, respectively. Some discussion is given in Section 2.7. The technical details of the proofs are given in Section A.

2.2 Setup and Literature Review

Let Y denote the response variable and (\mathbf{X}, \mathbf{Z}) the vector of covariates. The model of interest is the regression of Y on (\mathbf{X}, \mathbf{Z}) specified by

$$E(Y|\mathbf{X}, \mathbf{Z}) = g(\mathbf{X}, \mathbf{Z}; \beta_0), \quad (2.1)$$

where $g(\cdot)$ is a known link function continuously differentiable with respect to β , which is the regression parameter with true value β_0 . When data are fully observed, a typical way of estimating β_0 is to solve

$$\sum_{i=1}^n \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta) = \mathbf{0},$$

where $\mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \beta) = \mathbf{d}(\mathbf{X}, \mathbf{Z}; \beta)\epsilon(\beta)$, $\epsilon(\beta) = Y - g(\mathbf{X}, \mathbf{Z}; \beta)$ and $\mathbf{d}(\mathbf{X}, \mathbf{Z}; \beta)$ is a user-specified function of (\mathbf{X}, \mathbf{Z}) and may depend on β as well. One example is

$$\mathbf{d}(\mathbf{X}, \mathbf{Z}; \beta) = \frac{\partial g(\mathbf{X}, \mathbf{Z}; \beta)}{\partial \beta} \text{Var}(Y|\mathbf{X}, \mathbf{Z})^{-1},$$

which leads to a semiparametrically efficient estimator for β_0 under the regression model (2.1) (e.g., Tsiatis 2006).

We consider the case where Y and \mathbf{X} are fully observed but \mathbf{Z} is subject to missingness. Let R denote an indicator variable such that $R = 1$ if \mathbf{Z} is observed and $R = 0$ if \mathbf{Z} is missing. The observed data are n independent and identically distributed copies of $(Y, R\mathbf{Z}, \mathbf{X}, R)$. In this chapter we consider the MNAR mechanism where the missingness of \mathbf{Z} can depend on the possibly missing \mathbf{Z} but is conditionally independent of Y given \mathbf{Z} and \mathbf{X} ; i.e., $R \perp Y \mid (\mathbf{X}, \mathbf{Z})$. Such an MNAR mechanism is oftentimes more plausible than the MAR mechanism, especially when the response Y is measured at a later time point.

Under this setting, we essentially have $P(R = 1|Y, \mathbf{X}, \mathbf{Z}) = P(R = 1|\mathbf{X}, \mathbf{Z}) := \pi(\mathbf{X}, \mathbf{Z})$

and thus

$$f(Y|\mathbf{X}, \mathbf{Z}, R=1) = \frac{f(Y|\mathbf{X}, \mathbf{Z})\pi(\mathbf{X}, \mathbf{Z})}{\int f(Y|\mathbf{X}, \mathbf{Z})\pi(\mathbf{X}, \mathbf{Z})dY} = \frac{f(Y|\mathbf{X}, \mathbf{Z})\pi(\mathbf{X}, \mathbf{Z})}{\int f(Y|\mathbf{X}, \mathbf{Z})dY\pi(\mathbf{X}, \mathbf{Z})} = f(Y|\mathbf{X}, \mathbf{Z}).$$

The complete-case analysis by solving

$$\sum_{i=1}^n R_i \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) = \mathbf{0}$$

yields a consistent estimator for $\boldsymbol{\beta}_0$. However, the complete-case analysis does not use any information from the partially observed subjects and thus may not have the desired level of estimation efficiency.

To improve efficiency over the complete-case analysis, additional model assumptions other than (2.1) need to be made. Bartlett et al. (2014) assumed a logistic regression model $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ for $P(R=1|Y, \mathbf{X})$, where the parameter $\boldsymbol{\alpha}$ has true value $\boldsymbol{\alpha}_0$ such that $\pi(Y, \mathbf{X}; \boldsymbol{\alpha}_0) = P(R=1|Y, \mathbf{X})$. Since both Y and \mathbf{X} are fully observed, a consistent estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$ can be obtained by maximizing the binomial likelihood

$$\prod_{i=1}^n \pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})^{R_i} \{1 - \pi(Y_i, \mathbf{X}_i; \boldsymbol{\alpha})\}^{1-R_i}. \quad (2.2)$$

Bartlett et al. (2014) proposed the augmented complete-case (ACC) estimator $\hat{\boldsymbol{\beta}}_{\text{ACC}}$ for $\boldsymbol{\beta}_0$ by solving

$$\sum_{i=1}^n \{R_i \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) + \mathbf{V}(Y_i, \mathbf{X}_i, R_i; \boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})\} = \mathbf{0}, \quad (2.3)$$

where $\mathbf{V}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \{R - \pi(Y, \mathbf{X}; \boldsymbol{\alpha})\}\boldsymbol{\phi}(Y, \mathbf{X}; \boldsymbol{\beta})$ and $\boldsymbol{\phi}(Y, \mathbf{X}; \boldsymbol{\beta})$ is a user-specified function that has the same dimension as $\boldsymbol{\beta}$. They showed that the optimal $\boldsymbol{\phi}(Y, \mathbf{X}; \boldsymbol{\beta})$ that

leads to the smallest asymptotic variance of $\hat{\beta}_{\text{ACC}}$ is

$$\phi_{\text{opt}}(Y, \mathbf{X}; \beta) = -E\{U(Y, \mathbf{X}, \mathbf{Z}; \beta) \mid Y, \mathbf{X}, R = 1\}. \quad (2.4)$$

When a non-optimal $\phi(Y, \mathbf{X}; \beta)$ is used, however, although $\hat{\beta}_{\text{ACC}}$ is still consistent, it may lose efficiency compared to the complete-case analysis. In this case, Bartlett et al. (2014) proposed a modification to (2.3) so that the resulting estimator is at least as efficient as the complete-case analysis. But the implementation of this modification is difficult in general.

Noticing that both $RU(Y, \mathbf{X}, \mathbf{Z}; \beta)$ and $\mathbf{V}(Y, \mathbf{X}, R; \beta, \alpha)$ in (2.3) have mean zero when evaluated at β_0 and α_0 , Xie and Zhang (2017) considered the over-identified estimating function

$$\begin{pmatrix} RU(Y, \mathbf{X}, \mathbf{Z}; \beta) \\ \mathbf{V}(Y, \mathbf{X}, R; \beta, \hat{\alpha}) \end{pmatrix} \quad (2.5)$$

for β . They also considered combining this estimating function with the score function for α corresponding to (2.2) to form another over-identified estimating function

$$\begin{pmatrix} RU(Y, \mathbf{X}, \mathbf{Z}; \beta) \\ \mathbf{V}(Y, \mathbf{X}, R; \beta, \alpha) \\ \frac{R - \pi(Y, \mathbf{X}; \alpha)}{\pi(Y, \mathbf{X}; \alpha)\{1 - \pi(Y, \mathbf{X}; \alpha)\}} \frac{\partial \pi(Y, \mathbf{X}; \alpha)}{\partial \alpha} \end{pmatrix} \quad (2.6)$$

for (β, α) . Xie and Zhang (2017) proposed to use the empirical likelihood method (Qin and Lawless 1994) to estimate β_0 based on the estimating functions in (2.5) or (2.6). They showed that, when $\phi_{\text{opt}}(Y, \mathbf{X}; \beta)$ is used, estimators based on both (2.5) and (2.6) are asymptotically equivalent to the ACC estimator. When a non-optimal $\phi(Y, \mathbf{X}; \beta)$ is used, the estimator based on (2.6) is at least as efficient as both the complete-case analysis and the estimator based on (2.5), but the estimator based on (2.5) may be less efficient than the complete-case analysis. Refer to Xie and Zhang (2017) for a more detailed efficiency comparison.

2.3 A General Estimation Framework

The methods in Bartlett et al. (2014) and Xie and Zhang (2017) assume a correct model for $P(R = 1|Y, \mathbf{X})$ in order to improve efficiency over the complete-case analysis. It is possible to achieve the same goal by assuming models for quantities other than $P(R = 1|Y, \mathbf{X})$. We propose a general estimation framework that can accommodate different modelling strategies and thus covers the methods in Bartlett et al. (2014) and Xie and Zhang (2017) as special cases.

In general, let $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ denote a set of estimating functions for $\boldsymbol{\beta}$, which depend on the fully observed variables, Y , \mathbf{X} and R , and some nuisance parameter $\boldsymbol{\theta}$ that is introduced when modeling quantities beyond (2.1). Combining $RU(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ and $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$, we have an over-identified set of estimating functions for $\boldsymbol{\beta}$. To estimate $\boldsymbol{\beta}_0$, we take the empirical likelihood method and define the estimator $\hat{\boldsymbol{\beta}}_{\text{EL}}$ through

$$\begin{aligned} \max_{p_i, \boldsymbol{\beta}, \boldsymbol{\theta}} \prod_{i=1}^n p_i \quad \text{subject to} \\ p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \begin{pmatrix} R_i U(Y_i, \mathbf{Z}_i, \mathbf{X}_i; \boldsymbol{\beta}) \\ \mathbf{h}(Y_i, \mathbf{X}_i, R_i; \boldsymbol{\beta}, \boldsymbol{\theta}) \end{pmatrix} = \mathbf{0}. \end{aligned} \quad (2.7)$$

Here we require the dimension of $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ be larger than the dimension of $\boldsymbol{\theta}$. A discussion on this point is given after Theorem 1 below.

Based on the results in Qin and Lawless (1994), we have the following theorem regarding the consistency and the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\text{EL}}$. The derivation is given in Section A.

Theorem 1. *If $E\{\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\} = \mathbf{0}$ for a unique $\boldsymbol{\theta}_0$, then under the regularity conditions as in Qin and Lawless (1994), $\hat{\boldsymbol{\beta}}_{\text{EL}}$ is consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{EL}} - \boldsymbol{\beta}_0)$ has an asymptotic*

normal distribution with mean zero and variance

$$\{E(RU_{\beta}^T)E(RUU^T)^{-1}E(RU_{\beta}) + \mathbf{A}\mathbf{B}\mathbf{A}^T\}^{-1}, \quad (2.8)$$

where $\mathbf{U} = \mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \beta_0)$, $\mathbf{U}_{\beta} = \partial \mathbf{U}(Y, \mathbf{X}, \mathbf{Z}; \beta_0) / \partial \beta$, $\mathbf{h} = \mathbf{h}(Y, \mathbf{X}, R; \beta_0, \theta_0)$, $\mathbf{h}_{\beta} = \partial \mathbf{h}(Y, \mathbf{X}, R; \beta_0, \theta_0) / \partial \beta$, $\mathbf{h}_{\theta} = \partial \mathbf{h}(Y, \mathbf{X}, R; \beta_0, \theta_0) / \partial \theta$,

$$\begin{aligned} \mathbf{A} &= \{E(RU_{\beta}^T)E(RUU^T)^{-1}E(RU\mathbf{h}^T) - E(\mathbf{h}_{\beta}^T)\} \{E(\mathbf{h}\mathbf{h}^T) - E(R\mathbf{h}U^T)E(RUU^T)^{-1}E(RU\mathbf{h}^T)\}^{-1}, \\ \mathbf{B} &= E(\mathbf{h}\mathbf{h}^T) - E(R\mathbf{h}U^T)E(RUU^T)^{-1}E(RU\mathbf{h}^T) \\ &\quad - E(\mathbf{h}_{\theta}) \left(E(\mathbf{h}_{\theta}^T) \{E(\mathbf{h}\mathbf{h}^T) - E(R\mathbf{h}U^T)E(RUU^T)^{-1}E(RU\mathbf{h}^T)\}^{-1} E(\mathbf{h}_{\theta}) \right)^{-1} E(\mathbf{h}_{\theta}^T). \end{aligned}$$

From Lemma 2 in Section A, \mathbf{B} is positive semi-definite and so is $\mathbf{A}\mathbf{B}\mathbf{A}^T$, therefore the asymptotic variance of $\hat{\beta}_{\text{EL}}$ is no larger than that of the complete-case analysis, which is

$$\{E(RU_{\beta}^T)E(RUU^T)^{-1}E(RU_{\beta})\}^{-1}.$$

It is crucial to ensure that the dimension of $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ is larger than the dimension of θ . Intuitively, only in this case does $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ provide extra information for the estimation of β_0 in addition to the information needed for estimating θ_0 . Mathematically, if the dimension of $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ is no larger than the dimension of θ , the constrained maximization (2.7) simply leads to $\hat{p}_i = 1/n$ and $\hat{\beta}_{\text{EL}}$ being the complete-case analysis estimator.

When assuming a correct model $\pi(Y, \mathbf{X}; \alpha)$ for $P(R = 1|Y, \mathbf{X})$, this general framework covers the methods in Bartlett et al. (2014) and Xie and Zhang (2017) by taking

$\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ to be

$$\begin{pmatrix} \mathbf{V}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \frac{R - \pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})\{1 - \pi(\boldsymbol{\alpha})\}} \frac{\partial \pi(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \end{pmatrix}$$

and $\boldsymbol{\theta}$ to be $\boldsymbol{\alpha}$.

Furthermore, this general framework allows the possibility of modeling quantities different from $P(R = 1|Y, \mathbf{X})$ to improve efficiency over the complete-case analysis. A straightforward example is to model $E(Y|\mathbf{X})$. For instance, assuming a model $E(Y|\mathbf{X}; \boldsymbol{\gamma}) = \mu(\gamma_c + \boldsymbol{\gamma}_X^T \mathbf{X})$ with a known link function $\mu(\cdot)$ and unknown parameter $\boldsymbol{\gamma}$, we may take $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ to be $\mathbf{d}(\mathbf{X})\{Y - \mu(\gamma_c + \boldsymbol{\gamma}_X^T \mathbf{X})\}$ and $\boldsymbol{\theta}$ to be $\boldsymbol{\gamma}$, where $\mathbf{d}(\mathbf{X})$ is a user-specified vector function of \mathbf{X} with dimension larger than the dimension of $\boldsymbol{\gamma}$. When this model is correctly specified in the sense that $E(Y|\mathbf{X}; \boldsymbol{\gamma}_0) = E(Y|\mathbf{X})$ for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, Theorem 1 guarantees that $\hat{\boldsymbol{\beta}}_{\text{EL}}$ is more efficient than the complete-case analysis.

Another example is to model both $P(R = 1|Y, \mathbf{X})$ and $E(Y|\mathbf{X})$. In this case we take $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ to be

$$\begin{pmatrix} \mathbf{V}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \frac{R - \pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})\{1 - \pi(\boldsymbol{\alpha})\}} \frac{\partial \pi(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\ \mathbf{d}(\mathbf{X})\{Y - \mu(\gamma_c + \boldsymbol{\gamma}_X^T \mathbf{X})\} \end{pmatrix}$$

and $\boldsymbol{\theta}$ to be $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. Consistency and efficiency improvement over the complete-case analysis in this case requires both $P(R = 1|Y, \mathbf{X})$ and $E(Y|\mathbf{X})$ to be correctly modeled.

Model compatibility issues may arise when modelling additional quantities since we have already assumed a model of interest (2.1). For example, the model in (2.1) may impose some restrictions on how to model $E(Y|\mathbf{X})$. Thus, model selection and model checking techniques are often needed to reduce the chance of model incompatibility.

2.4 Choices of Quantities to Model

The efficiency improvement over the CC analysis implied by Theorem 1 is achieved by making model assumptions in addition to the model of interest in (2.1). Bartlett et al. (2014) and Xie and Zhang (2017) assumed a model for $P(R = 1|Y, \mathbf{X})$. Other model assumptions can be considered as well. Different assumptions involve different amounts of information, and thus lead to different efficiency improvements over the CC analysis. Although it is natural to ask what quantities should be modelled in order to have the most improvement, providing an answer is tremendously challenging, if not impossible, since even for the two cases of modelling $P(R = 1|Y, \mathbf{X})$ and $E(Y|\mathbf{X})$ there does not seem to be a direct efficiency comparison. Note, for example, the complex dependence of \mathbf{ABA}^T in (2.8) on β and $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$. Such a generally non-simplifiable dependence makes it almost impossible to find the “best” quantity to model. Compounding the problem is the incompatibility issue where in many settings there is no mathematically compatible model. In this case, models that agree closely with the observed data will presumably lead to estimates with small bias and efficiency improvement, but this needs to be investigated using numerical studies.

To gain insight into what quantities should be modeled we consider a simpler situation by dropping the dependence of $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ on R , β and θ . In other words, we find the optimal estimating function $\mathbf{h}(Y, \mathbf{X})$ leading to the maximum efficiency gain with $E\{\mathbf{h}(Y, \mathbf{X})\} = \mathbf{0}$ under the true underlying distribution. From Theorem 1, with $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ replaced by $\mathbf{h}(Y, \mathbf{X})$, the asymptotic variance in (2.8) becomes $[E(R\mathbf{U}_\beta^T)\{\text{Var}(\text{Resid}(R\mathbf{U}, \mathbf{h}))\}^{-1}E(R\mathbf{U}_\beta)]^{-1}$, where $\text{Resid}(R\mathbf{U}, \mathbf{h}) = R\mathbf{U} - E(R\mathbf{U}\mathbf{h}^T)E(\mathbf{h}\mathbf{h}^T)^{-1}\mathbf{h}$ is the residual of the projection of $R\mathbf{U}$ on the linear space spanned by \mathbf{h} . Due to this special structure, simple algebra shows that the optimal $\mathbf{h}(Y, \mathbf{X})$ leading to the most efficiency improvement over the CC analysis is

given by

$$\begin{aligned}\mathbf{h}_{opt}(Y, \mathbf{X}) &= E\{RU(Y, \mathbf{X}, \mathbf{Z}; \beta_0)|Y, \mathbf{X}\} \\ &= P(R = 1|Y, \mathbf{X})E\{U(Y, \mathbf{X}, \mathbf{Z}; \beta_0)|Y, \mathbf{X}, R = 1\}.\end{aligned}$$

However, $\mathbf{h}_{opt}(Y, \mathbf{X})$ is not directly applicable due to its dependence on the unknown underlying data distribution. First, it depends on the data distribution through the unknown β_0 . To overcome this, we consider the estimating function $P(R = 1|Y, \mathbf{X})E\{U(Y, \mathbf{X}, \mathbf{Z}; \beta)|Y, \mathbf{X}, R = 1\}$ instead of $\mathbf{h}_{opt}(Y, \mathbf{X})$. Second, $\mathbf{h}_{opt}(Y, \mathbf{X})$ depends on the data distribution through the unknown $P(R = 1|Y, \mathbf{X})$ and $f(\mathbf{Z}|Y, \mathbf{X}, R = 1)$. To overcome this, we assume models $\pi(Y, \mathbf{X}; \alpha) = P(R = 1|Y, \mathbf{X}; \alpha)$ and $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ that depend on nuisance parameters α and γ . Based on these considerations, the auxiliary estimating function we suggest is

$$\mathbf{h}_{use}(Y, \mathbf{X}; \beta, \theta) = \pi(Y, \mathbf{X}; \alpha)E\{U(Y, \mathbf{X}, \mathbf{Z}; \beta)|Y, \mathbf{X}, R = 1; \gamma\},$$

where $\theta = (\alpha, \gamma)$ and $E\{U(Y, \mathbf{X}, \mathbf{Z}; \beta)|Y, \mathbf{X}, R = 1; \gamma\}$ is taken under the model $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$. It is easy to verify that $E\{\mathbf{h}_{use}(Y, \mathbf{X}; \beta_0, \theta_0)\} = \mathbf{0}$, where $\theta_0 = (\alpha_0, \gamma_0)$ and γ_0 is the true value of γ such that $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma_0) = f(\mathbf{Z}|Y, \mathbf{X}, R = 1)$. Based on reasons given below Theorem 1, we consider estimating α_0 and γ_0 jointly with β_0 . This consideration leads to replacing $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ in (2.7) by

$$\begin{pmatrix} \mathbf{h}_{use}(Y, \mathbf{X}; \beta, \theta) \\ \frac{R - \pi(Y, \mathbf{X}; \alpha)}{\pi(Y, \mathbf{X}; \alpha)\{1 - \pi(Y, \mathbf{X}; \alpha)\}} \frac{\partial \pi(Y, \mathbf{X}; \alpha)}{\partial \alpha} \\ RS(Y, \mathbf{X}, \mathbf{Z}; \gamma) \end{pmatrix}, \quad (2.9)$$

where the second component is the score function corresponding to (2.2) for estimating α_0 and $S(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ is a user-specified estimating function for estimating γ_0 such that

$E\{RS(Y, \mathbf{X}, \mathbf{Z}; \gamma_0)\} = \mathbf{0}$. For example, $\mathbf{S}(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ may be taken to be the score function corresponding to the model $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$.

Implementation based on (2.9) involves two model assumptions in addition to (2.1), one for $P(R = 1|Y, \mathbf{X})$ and one for $f(\mathbf{Z}|Y, \mathbf{X}, R = 1)$. Both models need to be correctly specified for the proposed estimator $\hat{\beta}_{EL}$ to be consistent. In comparison, the ACC estimator in Bartlett et al. (2014) treats the model $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ as a working model and its consistency only requires correct specification of $\pi(Y, \mathbf{X}; \alpha)$. However, when $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ is incorrectly specified, the ACC estimator may be less efficient than the CC estimator. Since the main objective is to improve efficiency over the CC estimator because it is already consistent, $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ still needs to be a “good” model for the ACC method, if not the “correct” one. On the other hand, as discussed at the end of Section 3, in the real world there is always some degree of misspecification for parametric models. Therefore, we think that (2.9) is also worth consideration in scenarios where the ACC method is expected to provide improvement over the CC analysis. Note that the model for $f(\mathbf{Z}|Y, \mathbf{X}, R = 1)$ is fitted based on the complete cases. Complications for specifying, fitting and checking this model may arise when \mathbf{Z} is multivariate, especially if it is a mix of continuous and discrete variables.

When the dimension of β is larger than that of γ , $RS(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ in (2.9) may be dropped in the implementation, because in this case $RU(Y, \mathbf{X}, \mathbf{Z}; \beta)$ combined with the first two components of (2.9) already provides a set of over-identified estimating functions for $(\beta_0, \alpha_0, \gamma_0)$. The benefit of dropping $RS(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ from (2.9) in this case is two-fold. First, the reduction of the total number of estimating functions may improve the numerical performance of the empirical likelihood method, especially when this number is large. Second and more importantly, it will substantially reduce the bias of $\hat{\beta}_{EL}$ when $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ is misspecified. The reason is that, when $f(\mathbf{Z}|Y, \mathbf{X}, R = 1; \gamma)$ is misspecified, $RS(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ provides “incorrect” information about the data distribution.

When this “incorrect” information is accommodated in calculating $\hat{\beta}_{EL}$ and $\hat{\gamma}$, it pulls $\hat{\beta}_{EL}$ away from the true value β_0 . Dropping $RS(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ removes this undesired impact. On the contrary, the ACC and Xie and Zhang’s (2017) methods still require $RS(Y, \mathbf{X}, \mathbf{Z}; \gamma)$ as the estimating function to estimate γ , and thus still make full use of this “incorrect” information. Because of this, our proposed estimator can become less biased than the ACC and Xie and Zhang’s (2017) when the model for $P(R = 1 \mid Y, \mathbf{X})$ is also misspecified. Simulation Study 2 in Section 5 provides numerical evidence supporting this intuition. This observation is of high importance because, in the real world, it is likely that models for $P(R = 1 \mid Y, \mathbf{X})$ and $f(\mathbf{Z} \mid Y, \mathbf{X}, R = 1)$ are both misspecified and none of the existing estimators is consistent. Thus a possibly smaller bias by our proposed method becomes highly desired.

We also note that $\mathbf{h}_{use}(Y, \mathbf{X}; \beta, \theta)$ does not have a rigorous theoretical justification, and (2.9) is not necessarily the “optimal” estimating function in theory. Although using (2.9) is guaranteed to improve efficiency over the CC analysis when corresponding models are correctly specified, there is not a direct efficiency comparison to the ACC method.

2.5 Simulation Studies

The implementation of the EL estimators is based on Lagrange multiplier methods and Newton-Raphson iterations. We write the Lagrangian function of the constrained optimization in (2.7) as

$$\mathcal{L} = \sum_{i=1}^N \log p_i + \sum_{i=1}^N \boldsymbol{\lambda}^T \mathbf{g}_i(\boldsymbol{\phi}) + \mu \left(\sum_{i=1}^N p_i - 1 \right). \quad (2.10)$$

The function \mathbf{g} stands for the combined estimating function with both RU and \mathbf{h} , and the parameter $\boldsymbol{\phi}$ includes all the parameters to be estimated. The Lagrange multiplier $\boldsymbol{\lambda}$ can

be found by solving $\sum_{i=1}^n [\mathbf{g}_i(\boldsymbol{\phi}) / \{1 + \boldsymbol{\lambda}^T \mathbf{g}_i(\boldsymbol{\phi})\}] = \mathbf{0}$, which is equivalent to minimizing the function

$$l(\boldsymbol{\phi}, \boldsymbol{\lambda}) = - \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^T \mathbf{g}_i(\boldsymbol{\phi})\}$$

with respect to $\boldsymbol{\lambda}$. The algorithm we use is as follows.

We start from an initial estimate of $\boldsymbol{\phi}$, say the CML estimate, which we denote as $\boldsymbol{\phi}^0$. In the k -th iteration ($k = 1, 2, \dots$) we have steps

Step $k.1$. Given $\boldsymbol{\phi}^{k-1}$, we use the optimization function “constrOptim” in R to solve for the $\boldsymbol{\lambda}^k$ which minimizes $l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda})$.

Step $k.2$. We compute the first order derivative vector (Jacobian) of $l(\boldsymbol{\phi}, \boldsymbol{\lambda}^k)$ with respect to $\boldsymbol{\phi}$ and evaluate it at $\boldsymbol{\phi}^{k-1}$ as

$$\frac{\partial l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)}{\partial \boldsymbol{\phi}} = - \sum_{i=1}^n \frac{\mathbf{g}_{\phi,i}(\boldsymbol{\phi}^{k-1})^T \boldsymbol{\lambda}^k}{1 + \boldsymbol{\lambda}^{kT}(\boldsymbol{\phi}^{k-1}) \mathbf{g}_i(\boldsymbol{\phi}^{k-1})} := J(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)$$

where the entries in matrices $U_{\phi,i}(\boldsymbol{\phi}) = \partial U_i(\boldsymbol{\phi}) / \partial \boldsymbol{\phi}^T$ are computed by numerical differentiation.

The second-order derivative (Hessian) matrix is approximated by

$$\frac{\partial^2 l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \approx - \frac{\partial l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)}{\partial \boldsymbol{\phi}} \frac{\partial l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)}{\partial \boldsymbol{\phi}^T} := H(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k).$$

Step $k.3$. We update the estimate of $\boldsymbol{\phi}$ with

$$\boldsymbol{\phi}^k = \boldsymbol{\phi}^{k-1} - s \{H(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)\}^{-1} J(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k),$$

where s is an adaptive steplength such that $l(\boldsymbol{\phi}^k, \boldsymbol{\lambda}^k) \geq l(\boldsymbol{\phi}^{k-1}, \boldsymbol{\lambda}^k)$, which can be found via

line search.

We repeat the above steps until convergence. This so-called saddle-point algorithm is widely recommended in the EL literature (e.g., Owen (2001), Han and Lawless (2019)). In our settings, it converges acceptably fast.

In this section, we present simulation results from two studies. The first is to compare with the ACC estimator as in (Bartlett et al. 2014), but we found it to be a very special construction in creating an outcome independent scenario. The second setup is a more natural proposal.

2.5.1 Study 1

This simulation study uses the setup in Bartlett et al. (2014). The data are generated as $R \sim \text{Bernoulli}(0.5)$ and

$$\begin{pmatrix} Y \\ X \\ Z \end{pmatrix} \Big| R \sim \mathcal{N} \left(\begin{pmatrix} 0.2R \\ 0 \\ R \end{pmatrix}, \begin{pmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix} \right),$$

and the observed data vector is (Y, X, RZ, R) . This data generating process implies that the missingness of Z is MNAR and $R \perp Y \mid (X, Z)$. In addition, it ensures that $P(R = 1 \mid Y, X)$ can be correctly modeled by a logistic regression. The conditional mean model of interest is $E(Y \mid X, Z) = \beta_c + \beta_X X + \beta_Z Z$ with $\beta_0 = (\beta_c, \beta_X, \beta_Z) = (0, 0.2, 0.2)$. This simulation takes

$$U(Y, X, Z; \beta) = (1, X, Z)^T (Y - \beta_c - \beta_X X - \beta_Z Z).$$

Following Bartlett et al. (2014), let $\text{logit}\{\pi(Y, X; \alpha)\} = \alpha_c + \alpha_Y Y + \alpha_X X$ be the correctly specified model for $P(R = 1 \mid Y, X)$, $f_1(Z \mid Y, X, R = 1; \gamma)$ the correctly specified model

$\mathcal{N}(\gamma_c + \gamma_Y Y + \gamma_X X, \gamma_\sigma^2)$ for $f(Z | Y, X, R = 1)$, and $f_2(Z | Y, X, R = 1; \gamma)$ the misspecified model $\mathcal{N}(\gamma_c + \gamma_Y Y^2 + \gamma_X X^2, \gamma_\sigma^2)$ for $f(Z | Y, X, R = 1)$. The two models for $f(Z | Y, X, R = 1)$ are used to calculate $\phi_{opt}(Y, X; \beta)$ in (2.4).

We present the performance of the following estimators.

1. The complete-case analysis estimator $\hat{\beta}_{CC}$.
2. Two ACC estimators $\hat{\beta}_{ACC-1}$ and $\hat{\beta}_{ACC-2}$, both of which use $\pi(Y, Z; \alpha)$, but $\hat{\beta}_{ACC-1}$ is based on $f_1(Z | Y, X, R = 1; \gamma)$ and $\hat{\beta}_{ACC-2}$ is based on $f_2(Z | Y, X, R = 1; \gamma)$.
3. Two ACC2 estimators $\hat{\beta}_{ACC2-1}$ and $\hat{\beta}_{ACC2-2}$ as proposed in Bartlett et al. (2014), based on the ACC estimators $\hat{\beta}_{ACC-1}$ and $\hat{\beta}_{ACC-2}$, respectively.
4. Two estimators from Xie and Zhang (2017) $\hat{\beta}_{XZ1-1}$ and $\hat{\beta}_{XZ1-2}$ based on (2.5), using the same models as those for $\hat{\beta}_{ACC-1}$ and $\hat{\beta}_{ACC-2}$, respectively.
5. Two estimators from Xie and Zhang (2017) $\hat{\beta}_{XZ2-1}$ and $\hat{\beta}_{XZ2-2}$ based on (2.6), using the same models as those for $\hat{\beta}_{ACC-1}$ and $\hat{\beta}_{ACC-2}$, respectively.
6. Two estimators $\hat{\beta}_{EL-1}$ and $\hat{\beta}_{EL-2}$ based on our proposed method with (2.9), using the same models as those for $\hat{\beta}_{ACC-1}$ and $\hat{\beta}_{ACC-2}$, respectively.

For $\hat{\beta}_{EL-1}$ and $\hat{\beta}_{EL-2}$, the $\mathbf{S}(Y, X, Z, \gamma)$ in (2.9) is taken to be

$$\begin{pmatrix} (1, Y, X)^T (Z - \gamma_c - \gamma_Y Y - \gamma_X X) \\ (Z - \gamma_c - \gamma_Y Y - \gamma_X X)^2 - \gamma_\sigma^2 \end{pmatrix}$$

and

$$\begin{pmatrix} (1, Y^2, X^2)^T (Z - \gamma_c - \gamma_Y Y^2 - \gamma_X X^2) \\ (Z - \gamma_c - \gamma_Y Y^2 - \gamma_X X^2)^2 - \gamma_\sigma^2 \end{pmatrix},$$

Table 2.1: Simulation results for Study 1.

$n = 400$	Bias (Empirical Standard Error)[Root Mean Square Error]								
	$\beta_c(\beta_{c0} = 0)$			$\beta_Z(\beta_{Z0} = 0.2)$			$\beta_X(\beta_{X0} = 0.2)$		
CC	0.001	(0.094)	[0.094]	0.001	(0.071)	[0.071]	0.001	(0.070)	[0.070]
ACC-1	0.001	(0.093)	[0.093]	0.002	(0.069)	[0.069]	0.001	(0.052)	[0.052]
ACC-2	0.003	(0.094)	[0.094]	-0.001	(0.072)	[0.072]	0.002	(0.053)	[0.053]
ACC2-1	0.005	(0.093)	[0.093]	-0.001	(0.069)	[0.069]	0.001	(0.053)	[0.053]
ACC2-2	0.007	(0.095)	[0.095]	-0.002	(0.070)	[0.070]	0.000	(0.054)	[0.054]
XZ1-1	0.003	(0.093)	[0.093]	-0.001	(0.069)	[0.069]	0.001	(0.053)	[0.053]
XZ1-2	0.001	(0.095)	[0.095]	0.000	(0.072)	[0.072]	0.001	(0.054)	[0.054]
XZ2-1	0.003	(0.093)	[0.093]	-0.001	(0.069)	[0.069]	0.001	(0.053)	[0.053]
XZ2-2	0.002	(0.095)	[0.095]	0.000	(0.072)	[0.072]	0.001	(0.054)	[0.054]
EL-1	0.001	(0.094)	[0.094]	0.001	(0.070)	[0.070]	0.001	(0.053)	[0.053]
EL-2	0.158	(0.072)	[0.173]	-0.151	(0.067)	[0.165]	0.045	(0.052)	[0.068]
$n = 1000$	Bias (Empirical Standard Error)[Root Mean Square Error]								
	$\beta_c(\beta_{c0} = 0)$			$\beta_Z(\beta_{Z0} = 0.2)$			$\beta_X(\beta_{X0} = 0.2)$		
CC	0.002	(0.064)	[0.064]	-0.001	(0.045)	[0.045]	-0.001	(0.043)	[0.043]
ACC-1	0.001	(0.063)	[0.063]	0.000	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
ACC-2	0.002	(0.065)	[0.065]	-0.001	(0.046)	[0.046]	0.000	(0.032)	[0.032]
ACC2-1	0.003	(0.064)	[0.064]	-0.001	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
ACC2-2	0.004	(0.064)	[0.064]	-0.002	(0.045)	[0.045]	-0.001	(0.033)	[0.033]
XZ1-1	0.002	(0.063)	[0.063]	-0.002	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
XZ1-2	0.001	(0.064)	[0.064]	-0.001	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
XZ2-1	0.003	(0.063)	[0.063]	-0.002	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
XZ2-2	0.002	(0.064)	[0.064]	-0.001	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
EL-1	0.001	(0.064)	[0.064]	0.000	(0.045)	[0.045]	-0.001	(0.032)	[0.032]
EL-2	0.155	(0.059)	[0.059]	-0.151	(0.051)	[0.159]	0.043	(0.033)	[0.054]

respectively.

Table 1 summarizes the simulation results based on 1000 replications. The root mean square error (RMSE) in the box bracket shows a combined measure of bias and standard error. It is seen that EL-1 based on correctly specified models performs equally well compared to the ACC and Xie and Zhang’s (2017) estimators using the same models, and all have improved efficiency over the CC estimator. Note that in this case the ACC estimating equation in (2.3) represents the best linear combination of $RU(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ and $\mathbf{V}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\alpha})$, and thus the corresponding ACC estimator has the maximum efficiency. It is also seen that EL-2 based on the misspecified model $f_2(Z | Y, X, R = 1; \boldsymbol{\gamma})$ is biased. However, we would like to point out that $f_2(Z | Y, X, R = 1; \boldsymbol{\gamma})$ is unlikely to be chosen as a model for $f(Z | Y, X, R = 1)$ in the real world. It includes quadratic effects of Y and X without any linear effects. The likelihood ratio test comparing models $f_1(Z | Y, X, R = 1; \boldsymbol{\gamma})$ and $f_2(Z | Y, X, R = 1; \boldsymbol{\gamma})$ to the normal linear regression with Y , X , YX , Y^2 and X^2 as regressors rejected the two models 60 and 985 times out of 1000 replications when $n = 400$, respectively, and these numbers became 52 and 1000 when $n = 1000$, showing that it would be extremely unlikely to choose $f_2(Z | Y, X, R = 1; \boldsymbol{\gamma})$ to model $f(Z | Y, X, R = 1)$. Therefore the bias of EL-2 in this scenario should not be interpreted exclusively as a sign against our proposed method but rather an indication of the need for a model consistent with the observed data.

2.5.2 Study 2

This study considers three covariates, $Z \sim \text{Exponential}(2)$, $W \sim N(0, 1)$, and $X | W \sim N(W, 1)$. These variables are set to mimic a real data example where Z is a variable of response to a sensitive question, which is non-negative-valued and skewed. Other fully observed baseline variables may be assumed to be jointly normal. Given the covariates, Y

is generated as $Y = \beta_c + \beta_X X + \beta_Z Z + \beta_W W + \epsilon$, where $\beta_0 = (\beta_c, \beta_X, \beta_Z, \beta_W) = (0, 1, 1, 1)$ and $\epsilon \sim N(0, 1)$ is independent of the covariates X , W and Z . The missingness of Z is generated as $P(R = 1 | Y, X, Z, W) = \text{expit}(1 - 0.5Z + X + W)$, under which about 50% of subjects have missing Z . The conditional mean model of interest is $E(Y|X, Z, W) = \beta_c + \beta_X X + \beta_Z Z + \beta_W W$, and this simulation takes

$$\mathbf{U}(Y, X, Z, W; \beta) = (1, X, Z, W)^T(Y - \beta_c - \beta_X X - \beta_Z Z - \beta_W W),$$

In this simulation setting, it is very challenging, if not impossible, to derive a correct model for $P(R = 1 | Y, X, W)$. We consider the logistic regression model

$$\text{logit}\{\pi(Y, X, W; \alpha)\} = \alpha_c + \alpha_Y Y + \alpha_X X + \alpha_W W,$$

which is misspecified. To assess the goodness-of-fit of this model (Model 1) to the observed data, we compare it to two more complex models; one is a logistic regression with all the main effects and two-way interactions of Y , X and W (Model 2), and the other is the generalized additive model (Hastie and Tibshirani 1990) with a logit link and all main effects of Y , X and W smoothed by 4th order splines (Model 3). Taking $n = 400$, out of 1000 replications, the likelihood ratio test rejected Model 1 54 times when comparing it to Model 2 and 143 times when comparing it to Model 3, and the numbers of rejections became 44 and 136 with $n = 1000$. Therefore, model $\pi(Y, X, W; \alpha)$ would not be rejected most of the time.

For $f(Z | Y, X, W, R = 1)$, it is also very difficult to specify the correct model. Instead, we consider the following three models: $f_1(Z | Y, X, W, R = 1; \gamma)$ is the truncated normal distribution $N(\gamma_c + \gamma_Y Y + \gamma_X X + \gamma_W W, \gamma_\sigma^2)I(X > 0)$, $f_2(Z | Y, X, W, R = 1; \gamma)$ is the truncated normal distribution $N(\gamma_c + \gamma_Y Y, \gamma_\sigma^2)I(X > 0)$, and $f_3(Z | Y, X, W, R = 1; \gamma)$ is the normal distribution $N(\gamma_c + \gamma_Y Y, \gamma_\sigma^2)$. These models are used to calculate $\phi_{opt}(Y, X, W; \beta)$

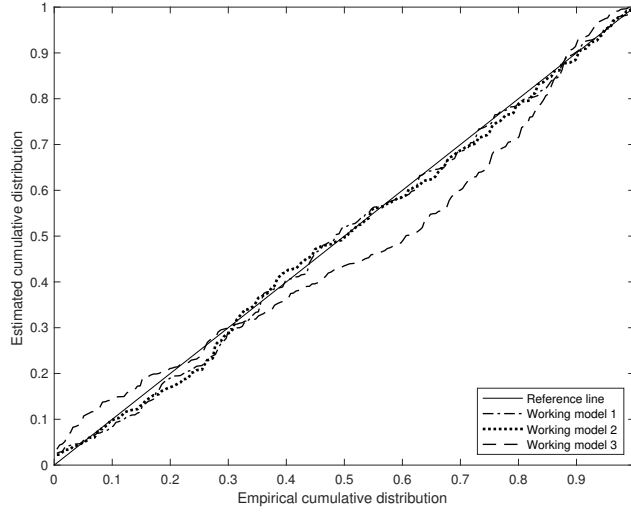


Figure 2.1: P-P plot for the three models for $f(Z | Y, X, W, R = 1)$ with $n = 400$

in (2.4). Figure 1 shows a typical P-P plot of these three models based on one simulation with $n = 400$. It clearly indicates that $f_3(Z | Y, X, W, R = 1; \gamma)$ is not a good model and is inferior to the other two. Samples with $n = 1000$ yield similar plots.

To further assess the goodness-of-fit of $f_1(Z | Y, X, W, R = 1; \gamma)$ and $f_2(Z | Y, X, W, R = 1; \gamma)$, we compare them to two more complex models; one is normal linear regression left truncated at 0 with all the main effects and two way interactions of Y , X and W (Model 4), and the other is normal linear regression left truncated at 0 with all the main and quadratic effects of Y , X and W (Model 5). Taking $n = 400$, out of 1000 replications, the likelihood ratio test rejected $f_1(Z | Y, X, W, R = 1; \gamma)$ 5 times and $f_2(Z | Y, X, W, R = 1; \gamma)$ 406 times when compared to Model 4 and 7 and 399 times when compared to Model 5. These numbers became 2, 981, 7 and 980 with $n = 1000$. In addition, the likelihood ratio test comparing $f_2(Z | Y, X, W, R = 1; \gamma)$ to $f_1(Z | Y, X, W, R = 1; \gamma)$ rejected the former 567 times with $n = 400$ and 994 times with $n = 1000$, out of 1000 replications. These tests suggest that $f_1(Z | Y, X, W, R = 1; \gamma)$ seems an adequate model whereas $f_2(Z | Y, X, W, R = 1; \gamma)$ is not.

Table 2 summarizes the simulation results based on 1000 replications. The CC, ACC,

Table 2.2: Simulation results for Study 2.

$n = 400$		Bias (Empirical Standard Error) [Root Mean Square Error]							
	$\beta_c(\beta_{c0} = 0)$		$\beta_z(\beta_{z0} = 1)$		$\beta_x(\beta_{x0} = 1)$		$\beta_w(\beta_{w0} = 1)$		
CC	0.004 (0.222) [0.222]		-0.007 (0.101) [0.101]		0.006 (0.149) [0.149]		-0.001 (0.206) [0.206]		
ACC-1	-0.168 (0.208) [0.267]		0.060 (0.096) [0.113]		0.036 (0.139) [0.144]		0.029 (0.190) [0.192]		
ACC-2	-0.161 (0.213) [0.267]		0.059 (0.099) [0.116]		0.036 (0.140) [0.145]		0.030 (0.191) [0.193]		
ACC-3	0.065 (0.243) [0.252]		0.003 (0.100) [0.100]		-0.050 (0.213) [0.219]		-0.058 (0.302) [0.308]		
ACC2-1	-0.142 (0.213) [0.257]		0.019 (0.101) [0.103]		0.044 (0.143) [0.150]		0.036 (0.193) [0.196]		
ACC2-2	-0.135 (0.215) [0.254]		0.015 (0.101) [0.102]		0.042 (0.143) [0.149]		0.034 (0.193) [0.196]		
ACC2-3	-0.043 (0.218) [0.222]		-0.005 (0.101) [0.101]		0.030 (0.142) [0.145]		0.024 (0.192) [0.194]		
XZ1-1	-0.225 (0.226) [0.319]		0.105 (0.110) [0.152]		0.036 (0.145) [0.149]		0.029 (0.198) [0.200]		
XZ1-2	-0.189 (0.217) [0.288]		0.088 (0.105) [0.137]		0.032 (0.142) [0.146]		0.026 (0.195) [0.197]		
XZ1-3	-0.150 (0.207) [0.256]		0.063 (0.099) [0.118]		0.029 (0.141) [0.144]		0.024 (0.193) [0.194]		
XZ2-1	-0.162 (0.232) [0.283]		0.043 (0.108) [0.116]		0.042 (0.146) [0.152]		0.036 (0.199) [0.202]		
XZ2-2	-0.170 (0.232) [0.288]		0.048 (0.107) [0.117]		0.042 (0.145) [0.151]		0.033 (0.198) [0.200]		
XZ2-3	-0.204 (0.217) [0.298]		0.061 (0.101) [0.119]		0.045 (0.145) [0.152]		0.040 (0.196) [0.200]		
EL-1	-0.177 (0.226) [0.287]		0.048 (0.108) [0.118]		0.048 (0.145) [0.153]		0.042 (0.198) [0.202]		
EL-2	0.310 (0.219) [0.379]		-0.153 (0.113) [0.190]		-0.048 (0.145) [0.153]		-0.055 (0.195) [0.202]		
EL-3	0.234 (0.201) [0.309]		-0.090 (0.096) [0.132]		-0.059 (0.134) [0.146]		-0.063 (0.185) [0.195]		
EL2-2	0.012 (0.223) [0.223]		-0.013 (0.101) [0.102]		0.005 (0.134) [0.134]		0.002 (0.188) [0.188]		
EL2-3	0.011 (0.223) [0.223]		-0.013 (0.101) [0.102]		0.006 (0.135) [0.135]		0.001 (0.189) [0.189]		
$n = 1000$		Bias (Empirical Standard Error) [Root Mean Square Error]							
	$\beta_c(\beta_{c0} = 0)$		$\beta_z(\beta_{z0} = 1)$		$\beta_x(\beta_{x0} = 1)$		$\beta_w(\beta_{w0} = 1)$		
CC	-0.003 (0.136) [0.136]		0.002 (0.061) [0.061]		-0.001 (0.094) [0.094]		-0.001 (0.130) [0.130]		
ACC-1	-0.170 (0.127) [0.212]		0.064 (0.059) [0.087]		0.030 (0.087) [0.092]		0.028 (0.119) [0.123]		
ACC-2	-0.166 (0.129) [0.210]		0.065 (0.060) [0.088]		0.032 (0.089) [0.094]		0.029 (0.120) [0.124]		
ACC-3	0.056 (0.149) [0.159]		0.012 (0.062) [0.063]		-0.058 (0.133) [0.145]		-0.055 (0.190) [0.197]		
ACC2-1	-0.150 (0.130) [0.199]		0.036 (0.061) [0.071]		0.034 (0.089) [0.095]		0.033 (0.121) [0.126]		
ACC2-2	-0.141 (0.131) [0.193]		0.032 (0.062) [0.070]		0.033 (0.089) [0.094]		0.031 (0.120) [0.124]		
ACC2-3	-0.050 (0.136) [0.145]		0.006 (0.062) [0.062]		0.024 (0.090) [0.093]		0.020 (0.124) [0.125]		
XZ1-1	-0.210 (0.134) [0.249]		0.095 (0.065) [0.115]		0.030 (0.091) [0.095]		0.030 (0.123) [0.126]		
XZ1-2	-0.189 (0.132) [0.230]		0.088 (0.064) [0.109]		0.027 (0.090) [0.094]		0.026 (0.121) [0.124]		
XZ1-3	-0.156 (0.127) [0.201]		0.067 (0.061) [0.091]		0.025 (0.090) [0.093]		0.023 (0.121) [0.123]		
XZ2-1	-0.149 (0.142) [0.206]		0.042 (0.071) [0.082]		0.035 (0.090) [0.097]		0.033 (0.122) [0.127]		
XZ2-2	-0.169 (0.143) [0.222]		0.058 (0.068) [0.090]		0.033 (0.090) [0.096]		0.031 (0.121) [0.125]		
XZ2-3	-0.198 (0.132) [0.238]		0.070 (0.063) [0.094]		0.037 (0.091) [0.099]		0.034 (0.123) [0.128]		
EL-1	-0.181 (0.136) [0.226]		0.062 (0.065) [0.090]		0.038 (0.092) [0.100]		0.037 (0.124) [0.129]		
EL-2	0.295 (0.158) [0.335]		-0.151 (0.081) [0.172]		-0.046 (0.094) [0.105]		-0.051 (0.129) [0.139]		
EL-3	0.211 (0.128) [0.247]		-0.074 (0.061) [0.096]		-0.059 (0.083) [0.102]		-0.064 (0.118) [0.134]		
EL2-2	0.008 (0.145) [0.145]		-0.005 (0.068) [0.068]		0.000 (0.086) [0.086]		-0.004 (0.117) [0.118]		
EL2-3	0.005 (0.140) [0.140]		-0.003 (0.063) [0.063]		0.000 (0.087) [0.087]		-0.004 (0.119) [0.119]		

ACC2, XZ1, XZ2, and EL estimators follow the same notation used in Study 1, now with three models $f_1(Z | Y, X, W, R = 1; \gamma)$, $f_2(Z | Y, X, W, R = 1; \gamma)$, and $f_3(Z | Y, X, W, R = 1; \gamma)$ considered. The $\mathbf{S}(Y, \mathbf{X}, \mathbf{Z}, \gamma)$ for estimating γ for all estimators is taken to be the score functions for these three models. Noting that $f_2(Z | Y, X, W, R = 1; \gamma)$ and $f_3(Z | Y, X, W, R = 1; \gamma)$ are clearly inadequate based on our model checking and γ has lower dimension than β for these two models, the two estimators EL2-2 and EL2-3 drop the $\mathbf{S}(Y, \mathbf{X}, \mathbf{Z}, \gamma)$ in (2.9), as discussed in Section 4. It is seen that, as expected, the ACC, ACC2, XZ1, XZ2, and EL estimators are all biased since neither $P(R = 1 | Y, X, W)$ nor $f(Z | Y, X, W, R = 1)$ is correctly modeled by any of the models under consideration, albeit the levels of bias vary somewhat. From Figure 2.1 we see that $f_3(Z | Y, X, W, R = 1; \gamma)$ is a clearly inadequate model compared to $f_1(Z | Y, X, W, R = 1; \gamma)$ and $f_2(Z | Y, X, W, R = 1; \gamma)$. However, the ACC-3 and ACC2-3 estimators based on it surprisingly have smaller bias than ACC-1 and ACC2-1 estimators based on a better model $f_1(Z | Y, X, W, R = 1; \gamma)$. Estimators EL2-2 and EL2-3 have very small bias, confirming the discussion in Section 4.

2.6 Data Application

As an application, we analyze the data collected in the year 2003-2004 from the US National Health and Nutrition Examination Survey (NHANES). NHANES is a program conducted by the Centers for Disease Control and Prevention to assess the health and nutritional status of both adults and children in the United States. We study the effect of average number of alcoholic drinks consumed per day on days when the subject drank alcohol (Z) on the systolic blood pressure (SBP) (mmHg) (Y), adjusting for age (in decade above 50) and body mass index (BMI) (kg/m^2) (\mathbf{X}). As pointed out in Little and Zhang (2011) and Bartlett et al. (2014), it is reasonable to assume that the SBP and BMI are missing completely at random, and thus in our analysis we only include the subjects with

these two variables fully observed. Among the $n = 2111$ subjects included in the analysis, 720 have missing values for alcohol consumption, and it is reasonable to assume this missingness depends on alcohol consumption itself but is independent of the SBP given alcohol consumption, age and BMI (Bartlett et al. 2014).

The model specifications follow Bartlett et al. (2014). The conditional mean model is

$$E(\text{SBP} \mid Z, \mathbf{X}) = \beta_c + \beta_1 \log(\text{no. of drinks} + 1) + \beta_2 \text{BMI} + \beta_3 \text{age} + \beta_4 \text{age}^2,$$

where SBP is centered at 125 mmHg and alcohol consumption is log transformed. For the missingness probability $P(R = 1 \mid Y, \mathbf{X})$, a logistic regression is assumed as

$$\text{logit}\{\pi(Y, \mathbf{X}; \boldsymbol{\alpha})\} = \alpha_c + \alpha_1 \text{age} + \alpha_2 \text{BMI} + \alpha_3 \text{SBP} + \alpha_4 \text{SBP}^2.$$

For $f(Z \mid Y, \mathbf{X}, R = 1)$, a negative binomial regression treating the number of drinks as the response is fitted, with all the linear and quadratic terms of age, BMI and SBP as regressors. The $\mathbf{U}(Y, \mathbf{X}, Z; \boldsymbol{\beta})$ is taken to be

$$\begin{pmatrix} 1 \\ \log(\text{no. of drinks} + 1) \\ \text{BMI} \\ \text{age} \\ \text{age}^2 \end{pmatrix} (\text{SBP} - \beta_c - \beta_1 \log(\text{no. of drinks} + 1) - \beta_2 \text{BMI} - \beta_3 \text{age} - \beta_4 \text{age}^2).$$

We calculate the complete-case estimator, the ACC estimator and our proposed empirical

Table 2.3: Analysis results for the NHANES data

	CC		ACC		EL	
	Estimate (SE)	<i>p</i> -value	Estimate (SE)	<i>p</i> -value	Estimate (SE)	<i>p</i> -value
Intercept	-1.929 (0.798)	0.015	-2.130 (0.741)	0.004	-1.921 (0.745)	0.010
Alcohol*	1.267 (0.583)	0.030	1.321 (0.598)	0.027	1.094 (0.550)	0.047
BMI	0.414 (0.080)	<0.001	0.388 (0.066)	<0.001	0.396 (0.062)	<0.001
Age	3.943 (0.261)	<0.001	3.888 (0.198)	<0.001	3.835 (0.227)	<0.001
Age ²	0.265 (0.143)	0.065	0.319 (0.104)	0.002	0.315 (0.107)	0.003

* log(number of drinks + 1)

likelihood estimator with $\mathbf{h}(Y, X, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ taken to be

$$\begin{pmatrix} \mathbf{h}_{use}(Y, \mathbf{X}, Z, R; \boldsymbol{\beta}, \boldsymbol{\theta}) \\ \frac{R - \pi(Y, \mathbf{X}; \boldsymbol{\alpha})}{\pi(Y, \mathbf{X}; \boldsymbol{\alpha})\{1 - \pi(Y, \mathbf{X}; \boldsymbol{\alpha})\}} \frac{\partial \pi(Y, \mathbf{X}; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\ R\mathbf{S}(Y, \mathbf{X}; \boldsymbol{\gamma}) \end{pmatrix},$$

where $\mathbf{S}(Y, \mathbf{X}; \boldsymbol{\gamma})$ is the score function for the regression coefficients of the negative binomial regression model for $f(Z | Y, \mathbf{X}, R = 1)$ (e.g. Lawless 1987). The relevant expectations under $f(Z | Y, \mathbf{X}, R = 1)$ needed in $\mathbf{h}_{use}(Y, \mathbf{X}, Z, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ are calculated by taking 200 random draws from the estimated negative binomial model. In our analysis we do not include the empirical likelihood estimator based on a model for $E(\text{SBP} | \mathbf{X})$ due to the consideration of compatibility between such a model and the model of interest $E(\text{SBP} | \mathbf{X}, Z)$.

Table 2.3 contains the results of our data analysis. All methods indicate that alcohol consumption is positively associated with increased SBP adjusting for the other covariates. The same conclusion can be made for BMI. Both ACC and the proposed method suggest a significant non-linear association between age and SBP, while the complete-case analysis fails to detect the significance.

2.7 Discussion

In our proposed method we jointly estimate the parameter of interest β and the nuisance parameter θ by solving estimating equations altogether using the empirical likelihood method. When the dimension of θ becomes large, the numerical performance by simultaneously solving all estimating equations may deteriorate. An alternative is to estimate part or all of θ separately by solving a lower dimensional estimating equation and then plug the estimated value into the rest of the estimating equations and solve for β using the empirical likelihood method. For example, α in $\pi(Y, \mathbf{X}; \alpha)$ and/or γ in $f(\mathbf{Z} | Y, \mathbf{X}, R = 1)$ can be separately estimated by maximizing (2.2) and $\prod_{i=1}^n f(\mathbf{Z}_i | Y_i, \mathbf{X}_i, R_i = 1)^{R_i}$, respectively, and then β can be estimated by the empirical likelihood method using the estimating function

$$\begin{pmatrix} RU(Y, \mathbf{X}, \mathbf{Z}; \beta) \\ \mathbf{h}_{use}(Y, \mathbf{X}, R; \beta, \hat{\theta}) \end{pmatrix}$$

with $\hat{\theta} = (\hat{\alpha}, \hat{\gamma})$ plugged in. The resulting estimator is called pseudo-empirical likelihood estimator in some literature and the asymptotic variance can be derived similarly (e.g., Qin, Zhang and Leung 2009). Based on our theoretical derivations, there is in general no clear efficiency comparison between this alternative method and the complete-case analysis, and thus the corresponding asymptotic results are not reported in this chapter. However, simulation studies not included here have shown that this alternative method does reduce the standard error compared to the complete-case analysis, and in some cases to a higher degree than estimating β and θ simultaneously.

The suggested $\mathbf{h}_{use}(Y, \mathbf{X}, R; \beta, \theta)$ in (2.9) requires models for both $P(R = 1 | Y, \mathbf{X})$ and $f(\mathbf{Z} | Y, \mathbf{X}, R = 1)$. This may seem undesirable compared to the ACC method which, in principle, only requires a model for $P(R = 1 | Y, \mathbf{X})$. However, the implementation of the ACC method would likely ask for a model for $f(\mathbf{Z} | Y, \mathbf{X}, R = 1)$ as well to estimate

$\phi_{opt}(Y, \mathbf{X}; \boldsymbol{\beta})$ in (2.4). Our simulation results have shown that our proposed estimator has less bias when such models are misspecified. Therefore, the proposed method indeed provides a useful alternative to ACC.

This work has been published as Che, Han & Lawless (2020).

Chapter 3

Empirical and Conditional Likelihoods for Two-Phase, Outcome-Dependent Samples

3.1 Introduction

Response-dependent samples arise in a variety of contexts; see for example, Qin (2017), where numerous examples and associated methodology are discussed. We consider two-phase studies with response-dependent samples that are designed to provide efficient estimation of covariate effects while limiting the cost of obtaining expensive covariate measurements (e.g. Song, Zhou & Kosorok 2009). Phase 1 data consisting of responses Y_i and covariates \mathbf{X}_i are available for a cohort of individuals $i = 1, \dots, N$ and in phase 2, values for expensive covariates \mathbf{Z}_i are obtained for a subset of n individuals. By appropriately basing phase 2 sample selection on the observed phase 1 responses and covariates, estimation efficiency can be increased. Such designs include generalized case-control and case-cohort designs used

with rare outcomes (e.g. Scott & Wild 2001, Borgan & Samuelsen 2014); other designs with response-dependent sampling are frequently used in genetic studies (e.g. Breslow et al. 2009; Barnett, Lee & Lin 2013; Lin, Zeng & Tang 2013) and many other areas.

We assume that triplets $(Y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \dots, N$ are iid and have joint probability density or mass function $f(Y|\mathbf{X}, \mathbf{Z})g(\mathbf{X}, \mathbf{Z})$. Our interest is in estimation of β in a regression model $f(Y|\mathbf{X}, \mathbf{Z}; \beta)$; the distribution $g(\mathbf{X}, \mathbf{Z})$ of the covariates is not of direct interest and we may wish to avoid modeling it (Scott and Wild 2011). Methods that avoid consideration of $g(\mathbf{X}, \mathbf{Z})$ include Horwitz-Thompson or inverse-probability-weighted (IPW) estimation and conditional maximum likelihood, e.g. see Lawless, Kalbfleisch & Wild (1999). Both approaches use only the data for phase 2 individuals but require specification of phase 2 selection probabilities. We denote $R_i = \mathbf{I}(\text{individual } i \text{ is selected for phase 2}), i = 1, \dots, N$ and let $\pi(y, \mathbf{x}, ; \alpha) = P(R_i = 1 | Y_i = y, \mathbf{X}_i = \mathbf{x})$. The phase 2 sample is typically selected either by variable probability sampling (VPS), where the R_i are independent Bernoulli random variables with selection probabilities $\pi(Y_i, \mathbf{X}_i)$ or by basic stratified sampling (BSS), in which case the N individuals are stratified on the basis of their phase 1 variables Y, \mathbf{X} , and a simple random sample is drawn from each stratum for phase 2. In the former case the true value α_0 of α is specified by design, but using an estimate $\hat{\alpha}$ instead can increase efficiency (Robins, Rotnitzky & Zhao 1994; Lawless, Kalbfleisch & Wild 1999; Scott & Wild 2011). For simplicity we focus on VPS in the discussion, but the results also apply to BSS with minor modifications (Scott & Wild 2011).

Conditional maximum likelihood (CML), described in the next section, has been widely used when Y is categorical (Breslow & Cain 1988; Keogh & Cox 2014; Qin 2017; Scott & Wild 1986) but also with continuous responses (Barnett, Lee & Lin 2013; Huang & Lin 2007; Li et al. 2011) and with censored failure time outcomes (Shen et al. 2015). Semi-parametric maximum likelihood (ML) methods are also used; they do not require a selection model $\pi(Y, \mathbf{X}; \alpha)$ but unlike CML, the covariate distribution must be estimated (Zhang &

Rockette 2005, 2007; Zhao, Lawless & McLeish 2009; Song, Zhou & Kosorok 2009; Zeng & Lin 2014; Tao, Zeng & Lin 2017). Full maximum likelihood is problematic computationally when \mathbf{X} is continuous or high-dimensional, and when \mathbf{Z} is high-dimensional, and CML is especially attractive in such cases. In addition, CML can be used when some $\pi(Y, \mathbf{X})$ may be zero (Barnett, Lee & Lin 2013; Huang & Lin 2007), whereas IPW estimation cannot. Conditional maximum likelihood estimates are not semiparametric efficient in general, but Scott & Wild (2011) showed how to augment CML so that full efficiency is achieved in certain situations. This occurs only for special discrete response models, and they did not show that their SW estimator dominates the CML estimator in general situations. We prove here that this is indeed the case.

We make four new contributions in this chapter. First, we combine CML and empirical likelihood (EL) (Owen 2001; Qin 2017) to improve on the efficiency of CML. Second, we prove that the EL and SW estimators of β have the same asymptotic variance and that in general, both dominate CML. This is of practical as well as theoretical interest because the SW estimators are computationally easier to obtain than the EL estimators. Third, we show that using a value of α that is known by design is asymptotically equivalent to estimating it for the EL estimators; this reduces computation somewhat. Finally, we provide numerical comparisons of CML, SW and EL estimators with both discrete and continuous response, and in the latter case compare them with semiparametric maximum likelihood. Section 3.2 outlines the CML, SW and EL procedures, and derives and compares asymptotic variances. Section 3.3 provides numerical comparisons of estimators in finite samples, and Section 3.4 discusses an illustration in a genetic testing context. Section 3.5 has concluding remarks. Details associated with derivations in Section 2 are given in an Appendix B. Section B.4 also contains two simulation studies that complement those in the chapter, and a few additional technical details which are mentioned in the chapter.

3.2 Estimators and Asymptotic Variances

3.2.1 The Conditional Maximum Likelihood (CML) Estimator

We will briefly review the CML estimator and set some notation. We assume that the phase 2 sample is selected using variable probability sampling (VPS). The CML estimator for β is based on the distribution of responses for the phase 2 individuals, conditional on covariate values and the fact they were selected for phase 2. The conditional likelihood (CL) is then

$$L_c(\beta, \alpha) = \prod_{R_i=1} P(Y = y_i | \mathbf{X} = \mathbf{x}_i, \mathbf{Z} = \mathbf{z}_i, R_i = 1) = \prod_{R_i=1} f_c(y_i | \mathbf{x}_i, \mathbf{z}_i; \beta, \alpha)$$

where, when Y is continuous,

$$f_c(Y | \mathbf{X}, \mathbf{Z}; \beta, \alpha) = \frac{f(Y | \mathbf{X}, \mathbf{Z}; \beta) \pi(Y, \mathbf{X}; \alpha)}{\int f(y | \mathbf{X}, \mathbf{Z}; \beta) \pi(y, \mathbf{X}; \alpha) dy}$$

and the conditional log-likelihood is

$$l_c(\beta, \alpha) = \sum_{i=1}^N R_i \log f_{ci}(\beta, \alpha) \tag{3.1}$$

where $f_{ci}(\beta, \alpha) := f_c(Y_i | \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha)$. With $\alpha = \alpha_0$ known by design, we can solve $\partial l_c(\beta, \alpha_0) / \partial \beta = \sum_{i=1}^N \mathbf{S}_{1,i}(\beta, \alpha_0) = \mathbf{0}$, where

$$\mathbf{S}_{1,i}(\beta, \alpha) = \mathbf{S}_1(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i; \beta, \alpha) = R_i \partial \log f_{ci}(\beta, \alpha) / \partial \beta$$

; we term the resulting estimator $\widehat{\beta}_{\text{CML0}}$. The asymptotic covariance matrix of $\sqrt{N}(\widehat{\beta}_{\text{CML0}} - \beta_0)$ is then

$$\begin{aligned}\Sigma_0 &= E \left\{ -\frac{\partial \mathbf{S}_1(\beta_0, \alpha_0)}{\partial \beta} \right\}^{-1} \text{Cov} \{ \mathbf{S}_1(\beta_0, \alpha_0) \} E \left\{ -\frac{\partial \mathbf{S}_1(\beta_0, \alpha_0)}{\partial \beta} \right\}^{-T} \\ &:= \mathcal{I}_{11}^{-1} \mathcal{C}_{11} \mathcal{I}_{11}^{-1}.\end{aligned}$$

Expectations here and later are taken with respect to the true distribution of $(Y, \mathbf{X}, \mathbf{Z}, R)$, given by $f(Y|\mathbf{X}, \mathbf{Z}; \beta_0)$, $g(\mathbf{X}, \mathbf{Z})$ and $\pi(Y, \mathbf{X}; \alpha_0)$. Since the components of (3.1) are log-likelihoods, it is easy to see that $\mathcal{I}_{11} = \mathcal{C}_{11}$, so $\Sigma_0 = \mathcal{C}_{11}^{-1}$.

Efficiency can be increased by estimating α even when it is known by design (Lawless, Kalbfleisch & Wild 1999); this can be done by using the likelihood function for α based on the observed values of R_1, \dots, R_N . Let $\phi = (\beta^T, \alpha^T)^T$ with true value ϕ_0 ; then the conditional maximum likelihood (CML) estimator $\widehat{\beta}_{\text{CML}}$ and ML estimator $\widehat{\alpha}_{\text{ML}}$ solve the estimating equation

$$\sum_{i=1}^N \mathbf{S}_{C,i}(\phi) = \sum_{i=1}^N \mathbf{S}_C(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i; \phi) = \sum_{i=1}^N \begin{pmatrix} \mathbf{S}_{1,i}(\beta, \alpha) \\ \mathbf{S}_{\pi,i}(\alpha) \end{pmatrix} = \sum_{i=1}^N \begin{pmatrix} \frac{R_i \partial \log f_{ci}}{\partial \beta} \\ \frac{R_i - \pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \alpha} \end{pmatrix} = \mathbf{0} \quad (3.2)$$

where $\pi_i = \pi(Y_i, \mathbf{X}_i; \alpha)$, $f_{ci} = f_{ci}(\beta, \alpha)$ and $\mathbf{S}_{\pi,i}(\alpha) = \mathbf{S}_{\pi}(Y_i, \mathbf{X}_i, R_i; \alpha) = (R_i - \pi_i)/\{\pi_i(1 - \pi_i)\} \partial \pi_i / \partial \alpha$. Then $\sqrt{N}\{(\widehat{\beta}_{\text{CML}}, \widehat{\alpha}_{\text{ML}}) - (\beta_0, \alpha_0)\}$ is asymptotically normal with covariance matrix $\mathcal{I}_{\text{CML}}^{-1} \mathcal{C}_{\text{CML}} \mathcal{I}_{\text{CML}}^{-T}$, where

$$\mathcal{C}_{\text{CML}} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{1\pi} \\ \mathcal{C}_{1\pi}^T & \mathcal{C}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} E\{\mathbf{S}_1(\phi_0) \mathbf{S}_1^T(\phi_0)\} & E\{\mathbf{S}_1(\phi_0) \mathbf{S}_{\pi}^T(\phi_0)\} \\ E\{\mathbf{S}_{\pi}(\phi_0) \mathbf{S}_1^T(\phi_0)\} & E\{\mathbf{S}_{\pi}(\phi_0) \mathbf{S}_{\pi}^T(\phi_0)\} \end{bmatrix},$$

and

$$\mathcal{I}_{\text{CML}} = E \left\{ -\frac{\partial \mathbf{S}_C(\phi_0)}{\partial \phi^T} \right\} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{1\pi} \\ \mathbf{0} & \mathcal{I}_{\pi\pi} \end{bmatrix}.$$

Since

$$\mathcal{I}_{\text{CML}}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & -\mathcal{I}_{11}^{-1} \mathcal{I}_{1\pi} \mathcal{I}_{\pi\pi}^{-1} \\ \mathbf{0} & \mathcal{I}_{\pi\pi}^{-1} \end{bmatrix},$$

and $\mathcal{I}_{1\pi} = E\{-\partial \mathbf{S}_1(\phi_0)/\partial \alpha^T\} = E\{\mathbf{S}_1(\phi_0) \mathbf{S}_\pi^T(\phi_0)\} = \mathcal{C}_{1\pi}$, $\mathcal{I}_{11} = \mathcal{C}_{11}$, and $\mathcal{I}_{\pi\pi} = \mathcal{C}_{\pi\pi}$ (details are in Appendix Section B.2), we find that the asymptotic covariance matrix of $\sqrt{N}(\hat{\beta}_{\text{CML}} - \beta_0)$ is

$$\begin{aligned} \Sigma_{\text{CML}} &= \mathcal{I}_{11}^{-1} - \mathcal{I}_{11}^{-1} \mathcal{I}_{1\pi} \mathcal{I}_{\pi\pi}^{-1} \mathcal{I}_{1\pi}^T \mathcal{I}_{11}^{-1} \\ &= \mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1} \mathcal{C}_{1\pi} \mathcal{C}_{\pi\pi}^{-1} \mathcal{C}_{1\pi}^T \mathcal{C}_{11}^{-1} \leq \mathcal{C}_{11}^{-1} = \Sigma_0, \end{aligned} \quad (3.3)$$

consistent with the well-known result that using the true value of α does not reduce the asymptotic variance of $\hat{\beta}$. Here and subsequently $A \leq B$ for positive definite symmetric matrices A and B means that $B - A$ is positive semi-definite.

3.2.2 The Scott-Wild (SW) Estimator

Noting that $l_c(\beta, \alpha)$ also carries information about α , Scott and Wild (2011) proposed an estimator that uses the score function $\mathbf{S}_2(Y, \mathbf{X}, \mathbf{Z}, R; \beta, \alpha) = R \partial \log f_c(Y|\mathbf{X}, \mathbf{Z}; \beta, \alpha)/\partial \alpha$. This was based on noticing that when Y is binary and $\pi(Y, \mathbf{X}; \alpha)$ depends only on a finite partition of \mathbf{Z} , the semi-parametric efficient maximum likelihood (ML) estimator developed

in Scott and Wild (1997) satisfies the estimating equation

$$\begin{aligned} \sum_{i=1}^N \mathbf{S}_{E,i}(\phi) &= \sum_{i=1}^N \mathbf{S}_E(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i; \phi) = \sum_{i=1}^N \begin{pmatrix} \mathbf{S}_{1,i}(\beta, \alpha) \\ \mathbf{S}_{\pi,i}(\alpha) - \mathbf{S}_{2,i}(\beta, \alpha) \end{pmatrix} \\ &= \mathbf{0}. \end{aligned} \quad (3.4)$$

However, this estimator is not fully efficient in general regression settings, for example with continuous Y , and Scott and Wild (2011) did not establish that it dominates the CML estimator based on (3.2) in general settings. We now do this.

The asymptotic covariance matrix of the solution to the estimating equation (3.4), which we denote as $\hat{\phi}_{\text{SW}}$, is given by $\mathcal{I}_{\text{SW}}^{-T} \mathcal{C}_{\text{SW}} \mathcal{I}_{\text{SW}}^{-1}$, where

$$\begin{aligned} \mathcal{I}_{\text{SW}} &= \begin{bmatrix} -E\{\partial \mathbf{S}_1(\phi_0)/\partial \beta^T\} & -E\{\partial \mathbf{S}_1(\phi_0)/\partial \alpha^T\} \\ E\{\partial \mathbf{S}_2(\phi_0)/\partial \beta^T\} & -E\{\partial (\mathbf{S}_\pi(\phi_0) - \mathbf{S}_2(\phi_0))/\partial \alpha^T\} \end{bmatrix} := \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{1\pi} \\ -\mathcal{I}_{21} & \mathcal{I}_{\pi\pi} - \mathcal{I}_{2\pi} \end{bmatrix}, \\ \mathcal{C}_{\text{SW}} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{1\pi} - \mathcal{C}_{12} \\ \mathcal{C}_{1\pi}^T - \mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{2\pi}^T - \mathcal{C}_{2\pi} + \mathcal{C}_{22} \end{bmatrix}, \end{aligned}$$

with $\mathcal{C}_{ij} = E\{\mathbf{S}_i(\phi_0) \mathbf{S}_j^T(\phi_0)\}$ for $i, j \in \{1, 2, \pi\}$.

In Section B.1 we show that $\mathcal{I}_{1\pi} = \mathcal{C}_{1\pi} = \mathcal{C}_{12}$ and $\mathcal{I}_{2\pi} = \mathcal{C}_{2\pi} = \mathcal{C}_{22}$. Also, \mathcal{C}_{22} is symmetric, so $\mathcal{C}_{2\pi} = \mathcal{C}_{22}^T = \mathcal{C}_{2\pi}^T$. Thus the asymptotic covariance matrix of $\sqrt{N}(\hat{\phi}_{\text{SW}} - \phi_0)$ is

$$\begin{aligned} &\mathcal{I}_{\text{SW}}^{-T} \mathcal{C}_{\text{SW}} \mathcal{I}_{\text{SW}}^{-1} \\ &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ -\mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{11} & \mathbf{0} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & -\mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{\pi\pi} - \mathcal{C}_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1} \mathcal{C}_{12}^T\}^{-1} & \mathbf{0} \\ \mathbf{0} & \{\mathcal{C}_{12}^T \mathcal{C}_{11}^{-1} \mathcal{C}_{12} + \mathcal{C}_{\pi\pi} - \mathcal{C}_{22}\}^{-1} \end{bmatrix} \end{aligned}$$

and the asymptotic covariance matrix of $\sqrt{N}(\hat{\beta}_{\text{SW}} - \beta_0)$ is

$$\begin{aligned}\Sigma_{\text{SW}} &= \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T\}^{-1} \\ &= \{\mathcal{C}_{11} + \mathcal{C}_{1\pi}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{1\pi}^T\}^{-1} \\ &= \mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22} + \mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi})^{-1}\mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}.\end{aligned}$$

Interestingly, $\hat{\beta}_{\text{SW}}$ and $\hat{\alpha}_{\text{SW}}$ are asymptotically uncorrelated. Since $\sqrt{N}(\hat{\beta}_{\text{CML}} - \beta_0)$ has asymptotic covariance matrix $\mathcal{C}_{11}^{-1} - \mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi}\mathcal{C}_{\pi\pi}^{-1}\mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}$ from (3.3), whether SW will improve upon CML depends on $\mathcal{C}_R := \mathcal{C}_{22} - \mathcal{C}_{1\pi}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{1\pi} = \mathcal{C}_{22} - \mathcal{C}_{12}^T\mathcal{C}_{11}^{-1}\mathcal{C}_{12}$. This is the asymptotic variance of \mathbf{S}_2 given \mathbf{S}_1 and so it is positive semi-definite; SW is therefore guaranteed to be as efficient as CML. That is, $\Sigma_{\text{SW}} \leq \Sigma_{\text{CML}}$.

3.2.3 The Empirical Likelihood Estimator

The Scott-Wild approach uses three estimating functions $\mathbf{S}_1(\beta, \alpha)$, $\mathbf{S}_2(\beta, \alpha)$ and $\mathbf{S}_\pi(\alpha)$, and it is possible that some information is lost by using only the difference of \mathbf{S}_π and \mathbf{S}_2 . Empirical likelihood (EL) allows us to exploit more estimating functions than the number of parameters (Qin & Lawless 1994, Owen 2001; Qin 2017). For two-phase, response-dependent samples, Zhou et al. (2011) gives an example of applying empirical likelihood but their method still needs to model the covariate distribution; we avoid this. We note that another approach to combining estimating functions is generalized method of moments, or GMM (Hansen 1982; Newey & Smith 2004). It is known that GMM and EL estimators based on a specific set of estimating functions have the same asymptotic distributions, but EL has some higher order asymptotic advantages (Newey & Smith 2004).

Let the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ be p and q , respectively and let

$$\mathbf{U}(\boldsymbol{\phi}) = (\mathbf{S}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})^T, \mathbf{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})^T, \mathbf{S}_\pi(\boldsymbol{\alpha})^T)^T.$$

Then, EL involves maximizing the empirical likelihood $\prod_{i=1}^N p_i$ with respect to $\mathbf{p} = (p_1, \dots, p_N), \boldsymbol{\beta}, \boldsymbol{\alpha}$ subject to $p_i \geq 0$, $\sum_{i=1}^N p_i = 1$ and the $p + 2q$ constraints

$$\sum_{i=1}^n p_i \mathbf{U}_i(\boldsymbol{\phi}) = \sum_{i=1}^n p_i \begin{pmatrix} R_i \frac{\partial \log f_{ci}}{\partial \boldsymbol{\beta}} \\ R_i \frac{\partial \log f_{ci}}{\partial \boldsymbol{\alpha}} \\ \left(\frac{R_i}{\pi_i} - \frac{1-R_i}{1-\pi_i} \right) \frac{\partial \pi_i}{\partial \boldsymbol{\alpha}} \end{pmatrix} = \mathbf{0}. \quad (3.5)$$

This produces an empirical likelihood estimator of $\boldsymbol{\phi}$, which we denote as $\hat{\boldsymbol{\phi}}_{\text{EL}}$. By Qin and Lawless (1994), it is asymptotically normal, with covariance matrix

$$(\mathcal{J}^T \mathcal{C}^{-1} \mathcal{J})^{-1} = \left[E \left\{ \frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T} \right\}^T E \{ \mathbf{U}(\boldsymbol{\phi}_0) \mathbf{U}^T(\boldsymbol{\phi}_0) \}^{-1} E \left\{ \frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T} \right\} \right]^{-1},$$

and in the Appendix Section B.1 we show

$$\mathcal{C} = E \{ \mathbf{U}(\boldsymbol{\phi}_0) \mathbf{U}^T(\boldsymbol{\phi}_0) \} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{22} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22}^T & \mathcal{C}_{\pi\pi} \end{bmatrix},$$

$$\mathcal{J} = -E \left\{ \frac{\partial \mathbf{U}(\boldsymbol{\phi}_0)}{\partial \boldsymbol{\phi}^T} \right\} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} \end{bmatrix}.$$

We now show that $\hat{\boldsymbol{\beta}}_{\text{SW}}$ and $\hat{\boldsymbol{\beta}}_{\text{EL}}$ have the same asymptotic covariance matrix. We write

the inverse of the asymptotic covariance matrix of $\sqrt{N}(\hat{\phi}_{\text{EL}} - \phi_0)$ in block form as

$$\begin{bmatrix} V^{11} & V^{12} \\ V^{21} & V^{22} \end{bmatrix},$$

and note that the asymptotic covariance matrix of $\sqrt{N}(\hat{\beta}_{\text{EL}} - \beta_0)$ can then be expressed as

$$\Sigma_{\text{EL}} : = \{V^{11} - V^{12}(V^{22})^{-1}V^{21}\}^{-1}.$$

In the Appendix Section [B.1](#) we show that

$$V^{11} = \mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T$$

and $V^{12} = \mathbf{0}$. Therefore

$$\begin{aligned} \Sigma_{\text{EL}} &= \{V^{11} - V^{12}(V^{22})^{-1}V^{21}\}^{-1} \\ &= \{\mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T\}^{-1} = \Sigma_{\text{SW}}. \end{aligned}$$

Thus $\hat{\beta}_{\text{SW}}$ and $\hat{\beta}_{\text{EL}}$ have the same asymptotic covariance matrix. This is of practical as well as theoretical importance, since the SW estimators are computationally easier to obtain than the EL estimators.

We remark that since phase 2 sampling is by design, the true value of α_0 is known. Qin, Zhang & Leung (2009) found with estimating functions used in the augmented inverse probability weighted (AIPW) estimator (Robins et al. 1994) that efficiency could in some cases be improved further by using the known α_0 , unlike the situation with CML. Xie and Zhang (2017) found a similar effect in another missing data setting. We call this estimator $\hat{\beta}_{\text{EL0}}$ in the setting here. Another alternative estimator uses the maximum likelihood estimate

$\hat{\boldsymbol{\alpha}}_{\text{ML}}$ in (3.5); Qin, Zhang & Leung (2009) called this a pseudo empirical likelihood estimator. We denote the resulting estimator as $\hat{\boldsymbol{\beta}}_{\text{PEL}}$. In our setting these two estimators respectively involve replacing $\boldsymbol{\phi}$ in (3.5) with $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}_0^T)^T$ to give $\tilde{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)$, and with $(\boldsymbol{\beta}^T, \hat{\boldsymbol{\alpha}}_{\text{ML}}^T)^T$ to give $\tilde{U}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}_{\text{ML}})$. The asymptotic covariance matrices for $\hat{\boldsymbol{\beta}}_{\text{EL0}}$ and $\hat{\boldsymbol{\beta}}_{\text{PEL}}$, denoted as $\boldsymbol{\Sigma}_{\text{EL0}}$ and $\boldsymbol{\Sigma}_{\text{PEL}}$ respectively, can both be shown to equal $(\tilde{\mathcal{I}}^T \mathcal{C}^{-1} \tilde{\mathcal{I}})^{-1}$, where \mathcal{C} is given above and

$$\tilde{\mathcal{I}} = -E \left\{ \frac{\partial \tilde{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}^T} \right\} = \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{21} \\ \mathbf{0} \end{bmatrix}.$$

Similarly to the derivation above, we find $\boldsymbol{\Sigma}_{\text{EL0}}^{-1} = \boldsymbol{\Sigma}_{\text{PEL}}^{-1} = V^{11}$ and therefore $\boldsymbol{\Sigma}_{\text{EL0}} = \boldsymbol{\Sigma}_{\text{PEL}} = \boldsymbol{\Sigma}_{\text{EL}} = \boldsymbol{\Sigma}_{\text{SW}}$. Thus, in our setting all the empirical likelihood estimators of $\boldsymbol{\beta}$ have the same asymptotic variance. Another option is to use known $\boldsymbol{\alpha}_0$ only in $\mathbf{S}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and to estimate $\boldsymbol{\alpha}$; this again gives the same asymptotic variance.

The asymptotic variances for all estimators of $\boldsymbol{\beta}$ depend on the phase 2 selection model $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ that is used in the system of estimating functions. It has been shown that efficiency can be improved by using a “highly stratified” model that includes the actual sampling probability model but incorporates a finer stratification of (Y, \mathbf{X}) than was actually used for sampling in the study, (e.g., Lawless, Kalbfleisch & Wild 1999; Scott & Wild 2011). For example, if phase 2 sampling depended only on Y , we can do better by using a working model that involves both Y and \mathbf{X} than with one that involves only Y . This is implicit in the asymptotic variances given here, and numerical studies in the next section illustrate the efficiency gains in finite samples.

3.3 Simulation Studies

Through the simulations, our EL estimators are implemented similarly to that in Chapter 2 (Section 2.5). The algorithm is outlined in Section 2.5, except that we used “fmincon” in MATLAB for the inner loop.

We describe numerical comparisons of the CML, SW and EL estimators for two scenarios here. The first is similar to one in Scott & Wild (2011) and involves a binary response Y with two covariates X and Z . Study 2 involves a continuous response. Two additional related scenarios are considered in Section B.4, as described below.

3.3.1 Simulation Study 1

This study involves binary covariate Z and continuous covariate X , which are correlated. We consider a phase 1 sample of 10,000 subjects with data generated as follows. A continuous standard normal covariate X_i is first generated and then a Bernoulli covariate Z_i is generated with probability $P(Z_i = 1) = 0.2I(X_i < 0) + 0.5I(X_i \geq 0)$. We then generate the response Y_i using a logistic regression model; with $\text{expit}(u)$ denoting $e^u/(1 + e^u)$, it is

$$P(Y = 1|X, Z) = \text{expit}(\beta_c + \beta_X X + \beta_Z Z), \quad (3.6)$$

with $\beta_0 = (-2.8, 0.5, 1)$. This gives marginal probabilities for $Y = 0$ and $Y = 1$ approximately equal to 0.9 and 0.1. We use a Bernoulli VPS selection scheme for phase 2, according to the sampling model

$$P(R = 1|Y, X) = \pi(Y, X; \alpha) = \text{expit}(\alpha_c + \alpha_Y Y), \quad (3.7)$$

where we chose $\alpha_0 = (-4.1, 2.5)$. This gives marginal selection probabilities for $Y = 0$ and $Y = 1$ of approximately 0.0163 and 0.168. We can use these known selection probabilities in our estimating functions; we refer to the CML estimator in this case as CML0. Estimation efficiency can be increased by estimating selection probabilities using the model (3.7), and we denote estimates obtained using this model with the suffix est in Table 3.1. It is possible, however, to further increase efficiency by using a more highly stratified selection model that conditions on observed X values, similar to calibration or post-stratification in sampling contexts. We consider two such models, referred to with the suffixes sat1 and sat2 in Table 3.1. For sat1 we use a binary covariate $V = I(X > 0.5)$ and the model

$$P(R = 1|Y, V) = \pi_{sat1}(Y, V; \alpha) = \text{expit}(\alpha_c + \alpha_Y Y + \alpha_V V + \alpha_{YV} YV). \quad (3.8)$$

The sat2 model uses the continuous covariate X in a more highly stratified logistic regression model for phase 2 selection, namely

$$P(R = 1|Y, X) = \pi_{sat2}(Y, X; \alpha) = \text{expit}(\alpha_c + \alpha_Y Y + \alpha_X X + \alpha_{YX} YX). \quad (3.9)$$

Note that working models (8) and (9) both include the true phase 2 sampling model (3.7) as special cases.

We also considered pseudo empirical likelihood (PEL) estimators, where the α parameters in models (3.7), (3.8) or (3.9) are first estimated by maximum likelihood from $\mathbf{S}_\pi(\alpha) = 0$ and then fixed in the estimating function $\mathbf{U}(\phi) = \mathbf{U}(\beta, \hat{\alpha}_{ML})$. This EL procedure is slightly easier to implement since the estimating function $\mathbf{S}_\pi(\hat{\alpha}_{ML})$ equals zero.

We mention that in this example the estimating functions \mathbf{S}_1 and \mathbf{S}_2 are not linearly independent when the π_{sat1} model is used. Then $\dim(\beta) = 3$ and $\dim(\alpha) = 4$ so the dimension of $(\mathbf{S}_1^T, \mathbf{S}_2^T)^T$ is 7. However in Section B.3 we show that the actual rank of these

7 estimating equations is 4. Therefore we use here only the first element of \mathbf{S}_2 for the EL estimator in this case. This phenomenon is an example of the well known fact that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not identifiable from the conditional likelihood $l_c(\boldsymbol{\beta}, \boldsymbol{\alpha})$ alone in this setting.

In Table 3.1, we compare the performance of CML, SW and EL estimators based on 500 simulations, using each of the three π models (3.7-3.9) as well as the known selection probabilities, denoted for CML as CML0. The EL0 and PEL estimator with each π model are asymptotically equivalent to the corresponding EL estimator so are omitted; their finite sample performances are close to those of the EL estimators. We show empirical standard deviations and average standard errors for each estimator; standard errors are obtained by estimating asymptotic covariance matrices with sample covariance matrices evaluated at estimates of $\boldsymbol{\phi}$. These are labelled empirical and estimated standard error (SE) in the table and they are seen to be close in value. In this case, CML performs about as well as the EL and SW methods. A substantial efficiency gain for estimation of β_x , the coefficient for the covariate that is known for all individuals, occurs when the stratified selection model (3.8) is used instead of (3.7) for the EL and SW estimators. A big increase in efficiency for CML and small further increases in efficiency for EL and SW result from using the more highly stratified model (3.9).

In Section B.4, we present a simulation study based on the same data generating model as here, in which basic stratified sampling (BSS) instead of VPS is used for phase 2 selection. It shows results very similar to the ones in Table 3.1.

3.3.2 Simulation Study 2

In Study 2, we simulate a normal linear regression model, $f(Y|X, Z; \beta) = N(\beta_c + \beta_X X + \beta_Z Z, \sigma^2)$, with stratified phase 2 sampling based on Y . We generate Z from a standard normal distribution; an auxiliary random variable W is generated by $W|Z \sim \text{Exp}(0.2I(Z >$

Table 3.1: Simulation results for Study 1.

Method	Mean (Empirical SE)[Estimated SE]		
	$\beta_c (\beta_{c0} = -2.8)$	$\beta_x (\beta_{x0} = 0.5)$	$\beta_z (\beta_{z0} = 1)$
CML0	-2.814 (0.168)[0.165]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
CML-est	-2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
CML-sat1	-2.803 (0.116)[0.121]	0.509 (0.204)[0.200]	0.997 (0.245)[0.251]
CML-sat2	-2.803 (0.113)[0.119]	0.514 (0.123)[0.123]	0.997 (0.244)[0.251]
EL-est	-2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
EL-sat1	-2.803 (0.115)[0.121]	0.502 (0.133)[0.133]	0.997 (0.244)[0.250]
EL-sat2	-2.803 (0.113)[0.119]	0.515 (0.123)[0.122]	0.997 (0.245)[0.251]
SW-est	-2.802 (0.117)[0.123]	0.519 (0.258)[0.257]	0.997 (0.244)[0.252]
EL-sat1	-2.803 (0.115)[0.121]	0.509 (0.139)[0.129]	0.997 (0.245)[0.250]
EL-sat2	-2.803 (0.113)[0.119]	0.514 (0.123)[0.122]	0.997 (0.244)[0.251]

0) + 0.5I(Z ≤ 0)), and then X is a categorization of W based on break-points at 0.5 and 1.5, which are approximately the 1/3 and 2/3 quantiles of W, and coded as 0, 1, 2 in the first, second and third tertile. We take $\beta_0 = (0, 0.5, 0.5)$ and $\sigma = 1$. We consider a phase 2 sampling plan often used in genetic epidemiological studies, in which the lower and upper tails of the Y distribution are over-sampled. In phase 1, (Y, X) are observed for every individual and in phase 2, we use VPS with $P(R = 1|Y, X) = \pi(Y; \alpha) = \sum_{k=1}^K \alpha_k I(y \in S_k)$, where the α_k are specified probabilities and $S_k = (c_{k-1}, c_k]$ is the k-th stratum in a partition of the sample space for Y. We set K = 3, and $S_1 = (-\infty, -0.05]$, $S_2 = (-0.05, 0.9]$, $S_3 = (0.9, \infty)$; the values -0.05 and 0.9 are approximately the 1st and 3rd quartiles of Y. We consider frequently used extreme response sampling and set $\alpha_0 = (\alpha_1, \alpha_2, \alpha_3) = (0.2, 0, 0.3)$, so about 17% of a phase 1 sample is selected for phase 2.

In this scenario, the conditional density function for Y is

$$\begin{aligned}
 f_c(Y|X, Z; \beta, \alpha) &= \frac{\exp\{-(y - \beta_c - \beta_X X - \beta_Z Z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(y \in S_k)}{\int \exp\{-(Y - \beta_c - \beta_X X - \beta_Z Z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(Y \in S_k) dy} \\
 &= \frac{\exp\{-(Y - \beta_c - \beta_X X - \beta_Z Z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(y \in S_k)}{\sum_{k=1}^3 \alpha_k \{F((c_k - \beta_c - \beta_X X - \beta_Z Z)/\sigma) - F((c_{k-1} - \beta_c - \beta_X X - \beta_Z Z)/\sigma)\}},
 \end{aligned}$$

where F is the standard normal distribution function. The parameter $\boldsymbol{\alpha} = (\alpha_1, 0, \alpha_3)$, though known, can be estimated through estimating equation

$$\left(\frac{\sum_{Y_i \in S_1} R_i}{\sum_i I(Y_i \in S_1)} - \hat{\alpha}_1, \frac{\sum_{Y_i \in S_3} R_i}{\sum_i I(Y_i \in S_3)} - \hat{\alpha}_3 \right) = \mathbf{0}.$$

We also consider a more highly stratified model $\pi(Y, X)$; we define $\tilde{\boldsymbol{\alpha}} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{31}, \alpha_{32}, \alpha_{33})$, where $\alpha_{ij} = P(R = 1 | Y \in S_i, X = j - 1)$. We compare the performances of CML, EL and SW estimators in 500 simulations with phase 1 sample size $N = 1000$ in Table 3.2. Two semiparametric maximum likelihood (SML) estimators are also included, one based on the likelihood in Zhao et al. (2009), referred to as SML-ZLM, and one on that in Lawless (1997), referred to as SML-L. The SML-L estimator uses only the phase 1 stratum information for Y, X whereas the SML-ZLM estimator uses the exact values for Y, X . The results for EL and SW estimators are similar to those for studies 1 and 2; CML is slightly less efficient than these estimators. However, we see that the SW and EL estimators are less efficient than SML; standard errors for SW and EL regression coefficients are 12-18 percent larger than those for SML-ZLM.

In Section B.3 we present a second simulation study with a continuous response, but where X and Z are both continuous. Comparisons of CML, SW and EL are similar to those in Table 3.2, but with SW and EL showing slightly less improvement over CML.

3.4 Illustration

We consider an application to genetic association testing based on data from Genetic Analysis Workshop 17 (Almasy et al. 2011). These data use real genotype sequences from 697 individuals, obtained from the 1000 Genomes Project. The Workshop organizers simulated data on three continuous traits (outcomes) by using models based on the genotype

Table 3.2: Simulation results for Study 2.

Method	Mean (Empirical SE)[Estimated SE]			
	$\beta_c (\beta_{c0} = -2.8)$	$\beta_X (\beta_{X0} = 0.5)$	$\beta_Z (\beta_{Z0} = 0.5)$	$\sigma (\sigma_0 = 1)$
CML0	0.007 (0.118)[0.117]	0.496 (0.087)[0.090]	0.505 (0.075)[0.077]	0.991 (0.049)[0.050]
CML-est	0.006 (0.106)[0.109]	0.497 (0.087)[0.090]	0.506 (0.075)[0.076]	0.992 (0.049)[0.050]
CML-sat	0.003 (0.092)[0.099]	0.500 (0.071)[0.078]	0.507 (0.075)[0.076]	0.093 (0.048)[0.050]
EL-est	0.006 (0.104)[0.104]	0.497 (0.087)[0.090]	0.505 (0.075)[0.076]	0.992 (0.049)[0.050]
EL-sat	0.004 (0.088)[0.084]	0.500 (0.068)[0.067]	0.506 (0.076)[0.075]	0.992 (0.049)[0.049]
SW-est	0.006 (0.103)[0.104]	0.497 (0.087)[0.090]	0.506 (0.075)[0.076]	0.992 (0.049)[0.050]
SW-sat	0.002 (0.085)[0.084]	0.501 (0.067)[0.067]	0.507 (0.074)[0.075]	0.993 (0.048)[0.049]
SML-ZLM	0.002 (0.075)	0.497 (0.058)	0.505 (0.064)	1.004 (0.034)
SML-L	0.003 (0.079)	0.500 (0.064)	0.505 (0.067)	1.004 (0.039)

information on each individual along with their sex, age and smoking status. There has been much interest in the last decade on testing for association between rare genetic variants and specific traits or outcomes (e.g Barnett et al. 2013; Derkach, Lawless and Sun 2014). We consider the problem of testing for an association between the quantitative trait called Q1 and rare variants in gene FLT1 on chromosome 13 (Yilmaz and Bull 2011). For the purposes of illustration, we consider the 321 individuals of Chinese or Japanese origin; this reduces population heterogeneity and avoids stratification issues. There are 18 single nucleotide polymorphisms (SNPs) with minor allele frequency less than 0.05 in this sub-population, and from these we derive a rare variant score for each individual as the total number of minor alleles across the 18 SNPs. These scores ranged from 0 to 3 across the 321 individuals.

As in Yilmaz and Bull (2011) we mimic a study in which the cost of genotyping all individuals is prohibitive and suppose that a two-phase design is used, in which the trait Q1 (Y in our notation) and covariates X represented by age (in years, and standardized), sex (male =1, female = 0) and smoking status (yes = 1, no = 0) are known for all 321 individuals. A phase 2 sample is then selected and for it the rare variant score Z is obtained.

We stratify the phase 1 sample by the quartiles of Q1 (Y). This gives three strata $S_1 = (-\infty, -0.6213)$ $S_2 = (-0.6213, 0.7826)$, and $S_3 = (0.7826, \infty)$, where -0.6213 and 0.7826 are

Table 3.3: Regression coefficient estimates for the GAW 17 two-phase data

Method	Mean [Estimated SE]							
	Intercept	# of rare variants		sex		age		smoke
Full data	-0.2489 [0.1701]	0.3164	[0.0857]	0.0344	[0.1011]	0.2805	[0.0505]	0.6878 [0.1176]
CML0	-0.2722 [0.2376]	0.2126	[0.0912]	0.1070	[0.1401]	0.3269	[0.0740]	0.7272 [0.1912]
CML-est	-0.2837 [0.2348]	0.2126	[0.0912]	0.1070	[0.1401]	0.3275	[0.0735]	0.7291 [0.1912]
CML-sat	-0.2703 [0.2310]	0.2128	[0.0912]	0.1086	[0.1374]	0.3239	[0.0700]	0.6655 [0.1667]
SW-est	-0.2889 [0.2348]	0.2126	[0.0912]	0.1070	[0.1401]	0.3278	[0.0735]	0.7300 [0.1913]
SW-sat	-0.2635 [0.2260]	0.2143	[0.0910]	0.1091	[0.1349]	0.3259	[0.0693]	0.6683 [0.1609]

the first and third quartiles of Q1. We consider a phase 2 VPS design which samples only from the first and third strata, with $P(R = 1|Y \in S_1) = P(R = 1|Y \in S_3) = 0.625$, giving approximately 100 individuals for phase 2. The true model for phase 2 selection is therefore $P(R = 1|Y) = \pi_{est}(Y; \boldsymbol{\alpha}) = \alpha_1 I(Y < -0.6213) + \alpha_3 I(Y \geq 0.7826)$, with $\alpha_1 = \alpha_3 = 0.625$. We also consider a more highly stratified model $P(R = 1|Y, Z) = \pi_{sat}(Y, Z; \tilde{\boldsymbol{\alpha}})$ where the π model is stratified also by smoking status and age. More specifically, we consider four strata determined by the sets $T_1 = I(\text{smoke} = 1, \text{age} \geq -0.087)$, $T_2 = I(\text{smoke} = 0, \text{age} \geq -0.087)$, $T_3 = I(\text{smoke} = 1, \text{age} < -0.087)$, $T_4 = I(\text{smoke} = 0, \text{age} < -0.087)$, where -0.087 is the median of (adjusted) age. Thus $\tilde{\boldsymbol{\alpha}} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \alpha_{34})$ with $\alpha_{ij} = P(R = 1|(Y, Z) \in S_i \times T_j)$. The estimates of regression coefficients in the normal regression model $N(\beta_c + \beta_X X + \boldsymbol{\beta}'_Z Z)$ for Y are shown in Table 3.3. For comparison we also show the estimates based on the full data for all 321 individuals, which is available. We see that the most efficient CML, SW and EL estimates based on X data for only about 30 percent of individuals have standard errors that are only moderately larger than those for the estimates from the full data. As in the numerical studies, the highly stratified selection models give more precise estimates, but the reductions in standard errors are not large in this setting.

3.5 Concluding Remarks

Our results show that for response-dependent two-phase studies the Scott-Wild (2011) estimator of β is optimal in the class of estimators based on the estimating functions $\mathbf{S}_1(\beta, \alpha)$, $\mathbf{S}_2(\beta, \alpha)$ and $\mathbf{S}_\pi(\alpha)$ for general regression models. The SW estimator, which is computationally straightforward to obtain, has the same asymptotic efficiency as more computationally demanding empirical likelihood estimators. We note that EL estimates of α appear slightly more efficient than SW estimates, but this is of no practical importance for two-phase studies since α_0 is known by design.

As expected, the finite sample performances of SW and EL in simulation studies were very similar. We have demonstrated that using more highly stratified models for $\pi(Y, \mathbf{X})$ in estimating functions can greatly increase efficiency. We also found that CML is very efficient when a sufficiently highly stratified model for π is used, especially with a binary response variable. Semiparametric maximum likelihood (ML) gave significant, though not huge, gains in efficiency over SW and EL in Study 3, where the response is continuous. When feasible computationally, ML is thus advantageous. Moreover, it can handle situations where phase 2 sampling is based on residuals from a fitted phase 1 model for Y given X (e.g. Derkach, Lawless and Sun 2015; Tao et al. 2017); CML and the corresponding SW and EL estimators cannot do this. However, in spite of recent advances in (Tao, Zeng & Lin 2017) ML that use kernel- or sieve-based methods for estimation of covariate distributions (Zeng & Lin 2014), (Tao et al. 2017), CML, SW and EL estimation remain attractive, especially with discrete responses and more generally, when covariate distributions are complex.

Finally, we note that recent advances have also been made in the optimal design of two-phase studies when ML is used (Tao et al. 2017). The designs are presumably highly efficient when CML, SW or EL estimation are used, but it would be useful to explore this.

This work has been published as Che, Lawless & Han (2020).

Chapter 4

Improving the Efficiency for Estimation with Two-Phase, Outcome-Dependent Samples

4.1 Introduction

Following Chapter 3, we further study the two-phase designs, which are cost-efficient for the estimation of a regression model with expensive covariates. In Phase 1, one measures the outcome Y and less expensive, or easy-to-measure covariate (vector) \mathbf{X} for the entire cohort, or representative sample consisting of individuals $i = 1, \dots, n$, and then in Phase 2, a subsample is taken to have the expensive covariate (vector) \mathbf{Z} measured. It is widely accepted that when we select the Phase 2 sample according to the Y values measured in Phase 1, known as outcome-dependent sampling (ODS), we can substantially reduce the study cost, as well as greatly improve the estimation efficiency of β (Breslow et al. 2009; Lin et al. 2013). Examples of ODS include case-control and case-cohort studies (Keogh & Cox 2014; Borgan &

Samuelson 2013) for rare outcomes and two-phase studies stratified by a continuous outcome (Huang & Lin 2007; Tao et al. 2017).

Our primary objective is to estimate β in a model $f(Y|\mathbf{X}, \mathbf{Z}; \beta)$. We denote $R_i = I(i\text{-th individual is selected into Phase 2})$, and since the Phase 2 selection only depends on the variates' values from Phase 1, a selection model can be written as $\pi_i(\alpha) = \pi(Y_i, \mathbf{X}_i; \alpha) = P(R_i = 1|Y_i, \mathbf{X}_i; \alpha)$, indexed by a nuisance parameter α . The correct model of π and the true value α_0 of α is then known by design. Therefore, the observed samples are $(Y_i, \mathbf{X}_i, R_i \mathbf{Z}_i, R_i)_{i=1}^n$ which is iid and satisfies $R \perp \mathbf{Z} | Y, \mathbf{X}$. In the context of missing data, the expensive covariate \mathbf{Z} is MAR.

As discussed in Chapter 1, methods of estimation for two-phase ODS fall into the following categories: maximum likelihood (ML) which models the likelihood of all observed data; pseudo-likelihood estimators with some estimated parts of the score function; IPW estimators with extensions such as AIPW; and conditional maximum likelihood (CML). The CML estimator is particularly attractive in many situations. Firstly, compared to ML and some weighting methods relying on a correct conditional covariate distribution, CML does not need the covariate distribution $g(\mathbf{Z}|\mathbf{X})$ to be specified. Parametric modeling of $g(\mathbf{Z}|\mathbf{X})$ is prone to misspecification, and leads to bias in parametric ML (Pepe & Fleming 1991). Non-parametric $g(\mathbf{Z}|\mathbf{X})$ is thus preferred to alleviate this problem and yields a semiparametric ML method (Zhang & Rockette 2006; Tao et al. 2017). However, the implementation would be cumbersome or infeasible when \mathbf{X} is continuous; or when \mathbf{Z} has dimension higher than 1. Secondly, compared to other methods without modeling $g(\mathbf{Z}|\mathbf{X})$, such as IPW, CML is known to be more efficient (Lawless et al. 1999; Scott & Wild 2011), and allows more flexible two-phase designs. Specifically, it can be applied when some individuals have selection probability $\pi_i = 0$ into Phase 2 which prohibits the use of IPW. Zero selection probabilities are quite common in applications. For example, in genome-wide association studies, when sequencing of the whole genome is expensive, researchers often sample those with extremely

high or low values of phenotypes, and this sampling strategy is shown to be more powerful and efficient than simple random sampling (Chen & Li 2011; Li, Lewinger, Gauderman, Murcray & Conti 2011; Lin et al. 2013; Bjørnland, Bye, Ryeng, Wisløff & Langaas 2018). Empirical genetic study examples include Padmanabhan, Melander, Johnson, Di Blasio, Lee, Gentilini, Hastie, Menni, Monti, Delles et al. (2010) for a hypertension study, where only individuals with extremely high or very low blood pressures are sampled for sequencing; Wang, Edmondson, Li, Gao, Qasim, Devaney, Burnett, Waterworth, Mooser, Grant et al. (2011) sampled subjects with extremely high high-density lipoprotein cholesterol (HDL-C) and another set of subjects with low HDL-C levels to identify novel pathways regulating HDL-C levels. In certain scenarios, the optimal design may involve zero selection probabilities for some individuals (Tao, Zeng & Lin 2019).

Classic CML still has room to be improved as it does not fully utilize the information contained in the Phase 1 individuals which are not selected into Phase 2. Many attempts have been made to use this information, including building ML estimators (Lawless et al. 1999; Weaver & Zhou 2005; Zhao, Lawless & McLeish 2009); incorporating Phase 1 information by using augmentation of conditional likelihood (Rivera-Rodriguez et al. 2020); or modeling π through post-stratification (Scott & Wild 2011; Che, Lawless & Han 2020). However, ML cannot avoid the modelling of $g(\mathbf{Z}|\mathbf{X})$; modelling π may not make full use of the Phase 1 data; and the augmentation is restricted to categorical outcomes (Rivera-Rodriguez et al. 2020). A systematic framework for utilizing the Phase 1 data effectively which applies to a broad range of settings is highly desired.

We propose a general framework to improve upon the classic CML estimator. Our framework maintains the ability of CML to allow zero selection probability in the Phase 2 sampling for certain subjects. Meanwhile, modelling the covariate distribution $g(\mathbf{Z}|\mathbf{X})$ is not required, hence it is suitable for a much wider range of data where the estimation of $g(\mathbf{Z}|\mathbf{X})$ is difficult, such as continuous \mathbf{X} , or \mathbf{Z} with dimension higher than 1. The auxiliary information

we propose to use is use a model for $f(Y|\mathbf{X})$. When π_i 's are strictly greater than 0 for every individual, we do not need this model to be correct, which mitigates the concern of incompatibility between the auxiliary estimating equation and the primary model of interest. When some Phase 1 individuals have zero selection probability of entering Phase 2, a correct specification of the model for $f(Y|\mathbf{X})$ is needed to guarantee efficiency improvement. However, in either case, $f(Y|\mathbf{X})$ is typically easy to model, as Y is a scalar, and values of both Y and \mathbf{X} are available for the entire cohort. Model diagnosis techniques can be applied. Even though the model for $f(Y|\mathbf{X})$ is subject to misspecification, the misspecification should be mild with the appropriate model diagnosis.

The rest of this chapter is organized as follows. Section 4.2 presents the framework and asymptotic properties under two settings, one with all individuals having positive probabilities of entering Phase 2, the other with some individuals having zero probabilities of entering Phase 2. We present some simulation results in Section 4.3 as well as an illustration with real data in Section 4.4. Finally, we give discussions in Section 4.5

4.2 Theory and Methods

4.2.1 Data and Model Setup

As detailed in Chapter 1, one can write the density model as in (1.10), with corresponding CML estimator solving the estimating equation

$$\sum_{i=1}^n R_i \mathbf{s}_{c,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0} \quad (4.1)$$

where $\mathbf{s}_{c,\beta}$ is the score function wrt $\boldsymbol{\beta}$. For the nuisance parameter $\boldsymbol{\alpha}$ in $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$, we can either use known $\boldsymbol{\alpha}_0$ or an estimated $\hat{\boldsymbol{\alpha}}$ from classic estimation such as maximum likelihood.

We denote the former as CML0 and the latter as CML. CML has higher efficiency than CML0, and may be further improved using post-stratification on the π model. We refer to Chapter 3 for a detailed discussion on variations of these estimators.

To make good use of the Phase 1 data, we can fit a reduced model $E_{(Y, \mathbf{X})}\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta})\} = \mathbf{0}$ from the following estimating equation

$$\sum_{i=1}^n \{\mathbf{h}(Y_i, \mathbf{X}_i; \boldsymbol{\theta})\} = \mathbf{0}, \quad (4.2)$$

with estimated parameter $\hat{\boldsymbol{\theta}}$. A natural choice here is to specify a working model $f(Y|\mathbf{X}; \boldsymbol{\theta})$ for $f(Y|\mathbf{X})$ and let $\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta})$ be the score function. Later we can show that for the case that every individual has a strictly positive probability of entering Phase 2, $\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta})$ does not need to be restricted to a score function. We let $\boldsymbol{\theta}^*$ denote the probability limit of $\boldsymbol{\theta}$, which satisfies $E_{(Y, \mathbf{X})}\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}^*)\} = \mathbf{0}$. It is to be noted that any working model $f(Y|\mathbf{X}; \boldsymbol{\theta})$ can be checked given that both Y and \mathbf{X} are observed for all individuals.

We now discuss two different scenarios. The first is $\pi(Y, \mathbf{Z}) > 0$ for all individuals and the other is $\pi(Y, \mathbf{Z}) = 0$ for certain individuals. We propose similar sets of estimators under these two different assumptions and make efficiency comparisons within either set.

4.2.2 Positive Selection Probability

In a lot of applications of two-phase studies, though the Phase 2 selection is outcome-dependent, there is no individual completely ruled out for Phase 2 selection, thus $\pi(Y, \mathbf{X})$ is always positive. Since the correct form of $\pi(Y; \mathbf{X})$ is known, we note that

$$\begin{aligned} \mathbf{0} &= E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}^*)\} \\ &= \int \frac{h(Y, \mathbf{X}; \boldsymbol{\theta}^*)P(R=1)}{\pi(Y, \mathbf{X})} \left\{ \frac{\pi(Y, \mathbf{X})}{P(R=1)} f(Y, \mathbf{X}, \mathbf{Z}) \right\} dY d\mathbf{X} d\mathbf{Z} \end{aligned}$$

$$\begin{aligned}
&= P(R = 1)E \left\{ \frac{h(Y, \mathbf{X}; \boldsymbol{\theta}^*)}{\pi(Y, \mathbf{X})} \middle| R = 1 \right\} \\
&= P(R = 1)E \left[E \left\{ \frac{h(Y, \mathbf{X}; \boldsymbol{\theta}^*)}{\pi(Y, \mathbf{X})} \middle| \mathbf{X}, \mathbf{Z}, R = 1 \right\} \middle| R = 1 \right] \\
&= P(R = 1)E \{ \mathbf{u}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*) | R = 1 \},
\end{aligned}$$

where

$$\mathbf{u}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \frac{h(Y, \mathbf{X}; \boldsymbol{\theta})}{\pi(Y, \mathbf{X})} f_c(Y | \mathbf{X}, \mathbf{Z}, R = 1; \boldsymbol{\beta}) dY. \quad (4.3)$$

Since $P(R = 1)$ is a constant, the above equation implies

$$E_{(\mathbf{X}, \mathbf{Z} | R=1)} \{ \mathbf{u}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}^*) | R = 1 \} = \mathbf{0}, \quad (4.4)$$

and puts a constraint on the conditional distribution $F(\mathbf{X}, \mathbf{Z} | R = 1)$. To incorporate the information about $\boldsymbol{\beta}$ contained in this moment condition, we consider the empirical probabilities $p_i := dF(\mathbf{X}_i, \mathbf{Z}_i | R_i = 1)$, with support on the Phase 2 data $i = 1, \dots, m$. We get an estimator $\hat{\boldsymbol{\beta}}_{\text{EL0}}$ defined through

$$\begin{aligned}
&\max_{\boldsymbol{\beta}, p_1, \dots, p_m} \prod_{i=1}^m f_c(Y_i | \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) p_i \quad \text{subject to} \\
&p_i \geq 0, \quad \sum_{i=1}^m p_i = 1, \quad \text{and} \quad \sum_{i=1}^m p_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^*) = \mathbf{0}.
\end{aligned} \quad (4.5)$$

Such a formulation is similar to those in Qin (2000), Chatterjee, Chen, Maas & Carroll (2016) and Han & Lawless (2019) .

We note that $\boldsymbol{\theta}^*$ in (4.5) is usually unknown, thus needs an estimate. However, when the Phase 1 sample size is large enough, uncertainty in the estimation can be ignored, i.e., we can take $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ in (4.5). We denote the estimator by plugging $\boldsymbol{\theta}^*$ into (4.5)

as $\hat{\beta}_{\text{EL0-1}}$, and write $\mathbf{S} := E\{R\mathbf{s}_{c,\beta}(\beta_0, \theta^*)\mathbf{s}_{c,\beta}(\beta_0, \theta^*)^T\}$, $\mathbf{J} := E\{R\partial\mathbf{u}(\beta_0, \theta^*)/\partial\beta^T\}$, and $\mathbf{\Omega} := E\{R\mathbf{u}(\beta_0, \theta^*)\mathbf{u}(\beta_0, \theta^*)^T\}$, with all the expectations taken with respect to the joint distribution $(Y, \mathbf{X}, \mathbf{Z}, R)$. In the Appendix we show that

$$\sqrt{n}(\hat{\beta}_{\text{EL0-1}} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_1), \quad (4.6)$$

where the asymptotic variance $\mathbf{\Sigma}_1 = (\mathbf{S} + \mathbf{J}\mathbf{\Omega}^{-1}\mathbf{J}^T)^{-1}$, and is guaranteed to be less than or equal to the asymptotic variance of $\hat{\beta}_{\text{CML}}$, which we denote as $\mathbf{\Sigma}_0$ and is easily shown to equal \mathbf{S}^{-1} .

Han and Lawless (2019) finds that an alternative empirical likelihood estimator leads to the same asymptotic efficiency as the empirical likelihood estimator defined in the form of (4.5). We can show similar results in our scenario. Denote $p_i = dF(Y, \mathbf{X}, \mathbf{Z}|R = 1)$, we can show that the asymptotic variance $\mathbf{\Sigma}_1$ is the same as an empirical likelihood estimator $\hat{\beta}_{\text{EL1}}$ defined through

$$\begin{aligned} & \max_{\beta, p_1, \dots, p_n} \prod_{i=1}^n p_i \quad \text{subject to } p_i \geq 0, \sum_{i=1}^n p_i = 1, \\ & \text{and } \sum_{i=1}^m p_i \begin{pmatrix} R_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta^*) \\ \mathbf{s}_{c,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \beta) \end{pmatrix} = \mathbf{0}. \end{aligned} \quad (4.7)$$

The equivalence is also shown in Chapter C.

When the Phase 1 sample is not too large compared to the Phase 2 sample, the uncertainty in estimating θ needs to be accounted for to obtain a valid variance estimate. When $\hat{\theta}$ replaces the θ^* in (4.5) and (4.7), we can still derive the equivalence between two empirical likelihood estimators, $\hat{\beta}_{\text{EL0-2}}$ and $\hat{\beta}_{\text{EL2}}$. Let $\mathbf{U} := E\{R\mathbf{s}_{c,\beta}(\beta_0, \alpha_0, \theta_0)\mathbf{h}(\theta_0)^T\}$, $\mathbf{V} := E\{R\mathbf{u}(\beta_0, \alpha_0, \theta_0)\mathbf{h}(\theta_0)^T\}$, and $\mathbf{W} := E\{\mathbf{h}(\theta_0)\mathbf{h}(\theta_0)^T\}$. Denoting the asymptotic

variance of $\widehat{\beta}_{\text{EL0-2}}$ and $\widehat{\beta}_{\text{EL2}}$ as Σ_2 , we show in Chapter C that

$$\Sigma_2 = (S + J\Omega^{-1}J^T)^{-1} \{S - J\Omega^{-1}U - U^T\Omega^{-1}J^T + J\Omega^{-1}(\Omega - V - V^T + W)\Omega^{-1}J^T\} \\ \cdot (S + J\Omega^{-1}J^T)^{-1}.$$

All of the above estimators use the known $\pi(Y; \mathbf{X})$. However, it is shown in a lot of literature that modeling the known $\pi(Y, \mathbf{X})$ and using the modeled values instead of the known $\pi(Y, \mathbf{X})$ further improves estimation efficiency (e.g., Scott and Wild 1997, 2011; Lawless et al. 1999). This is also similar to the observations made in the missing data literature (e.g., Robins et al. 1994; Tsiatis 2007). Therefore, we also consider postulating a parametric model $\pi(Y, \mathbf{X}; \alpha)$ for $\pi(Y, \mathbf{X})$. The postulated $\pi(Y, \mathbf{X}; \alpha)$ needs to include the correct $\pi(Y, \mathbf{X})$ so that the parameter α has true value α_0 . One way to make use of the postulated $\pi(Y, \mathbf{X}; \alpha)$ model is to replace $\pi(Y, \mathbf{X})$ by $\pi(Y, \mathbf{X}; \widehat{\alpha}_{\text{MLE}})$ in the above estimating equations (4.5, 4.7), where $\widehat{\alpha}_{\text{MLE}}$ maximizes the likelihood

$$\prod_{i=1}^n \{\pi(Y, \mathbf{X}; \alpha)\}^{R_i} \{1 - \pi(Y, \mathbf{X}; \alpha)\}^{1-R_i}, \quad (4.8)$$

or equivalently, it solves the score function

$$s_\alpha(\alpha) = s_\alpha(Y, \mathbf{X}, R; \alpha) = \frac{R - \pi(Y, \mathbf{X}; \alpha)}{\pi(Y, \mathbf{X}; \alpha)\{1 - \pi(Y, \mathbf{X}; \alpha)\}} \frac{\partial \pi(Y, \mathbf{X}; \alpha)}{\partial \alpha}.$$

However, Qin et al. (2009) finds that in a similar MAR setting, the estimation efficiency of β also benefits from estimating α simultaneously by including the score function corresponding to (4.8) in the estimating functions. We can thus propose an estimator $\widehat{\beta}_{\text{EL3}}$ through

$$\max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \quad \text{subject to}$$

$$q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{c,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{s}_\alpha(Y_i, \mathbf{X}_i, R_i; \alpha) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}, \quad (4.9)$$

where $q_i = dF(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i)$ denotes the empirical distribution with support on the observed data. By Corollary 1 of Qin & Lawless (1994), the EL3 estimator is at least as efficient as EL2.

EL3 may be further improved as we note that its estimating functions are those of CML with an estimated α plus auxiliary functions $\mathbf{u}(\mathbf{X}, \mathbf{Z}; \beta, \alpha, \theta)$ and $\mathbf{h}(Y, \mathbf{Z}; \beta, \alpha, \theta)$. Scott and Wild (2011) noted that, with $\pi(Y, \mathbf{X})$ modeled by $\pi(Y, \mathbf{X}; \alpha)$, the $f_c(Y|\mathbf{X}, \mathbf{Z}, R = 1; \beta)$ in (3.1) also depends on α , and we rewrite it as $f_c(Y|\mathbf{X}, \mathbf{Z}, R = 1; \beta, \alpha)$ to make this dependence explicit. Thus the score function component $\mathbf{s}_{c,\alpha}(\beta, \alpha) = \partial \log\{f_c(Y|\mathbf{X}, \mathbf{Z}, R = 1; \beta, \alpha)\}/\partial \alpha$ also contains information about both β and α , similar to the score function component $\mathbf{s}_{c,\beta}(\beta, \alpha) = \partial \log\{f_c(Y|\mathbf{X}, \mathbf{Z}, R = 1; \beta, \alpha)\}/\partial \beta$, previously denoted as $\mathbf{s}_{c,\beta}(\beta)$ when no model for $\pi(Y, \mathbf{X})$ was considered. Therefore, the estimating function $\mathbf{s}_{c,\alpha}(\beta, \alpha)$ should also be accounted for, similarly to $\mathbf{s}_{c,\beta}(\beta, \alpha)$, to further improve efficiency. In Chapter 3, we have explored the asymptotic properties of adding $\mathbf{s}_{c,\alpha}(\beta, \alpha)$ by different ways. With $q_i = dF(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i)$, an estimator $\hat{\beta}_{\text{EL4}}$ can be defined with $\mathbf{s}_{c,\alpha}(\beta, \alpha)$ added as part of the set of estimating functions in (4.9):

$$\max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \text{ subject to}$$

$$q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{c,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{s}_{c,\alpha}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{s}_\alpha(Y_i, \mathbf{X}_i, R_i; \alpha) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}, \quad (4.10)$$

The derivation of $\widehat{\beta}_{\text{EL4}}$ is based on a joint maximization with respect to β , α and θ . It makes use of every piece of information available in the form of an estimating function, and thus giving the maximum efficiency we are able to achieve in estimating β . A potential issue with $\widehat{\beta}_{\text{EL4}}$ is that in some special cases, especially when the response Y is binary, there may exist collinearity among the constraints in (4.10) (see Section B.3). A transform first proposed in Scott and Wild (2011) is suggested in Chapter 3. Similar idea may be employed here, that we can combine the two components $\mathbf{s}_\alpha(Y, \mathbf{X}, R; \alpha)$ and $R\mathbf{s}_{c,\alpha}(Y, \mathbf{X}, \mathbf{Z}; \beta, \alpha)$ in (4.10) by taking their difference and use the difference as an estimating function. That is, we can consider another estimator $\widehat{\beta}_{\text{EL5}}$ defined through

$$\max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \text{ subject to } q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{c,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ \mathbf{s}_\alpha(Y_i, \mathbf{X}_i, R_i; \alpha) - R_i \mathbf{s}_{c,\alpha}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}, \quad (4.11)$$

with $q_i = dF(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i)$. In the absence of collinearity among the constraints, it is easy to see that $\widehat{\beta}_{\text{EL4}}$ has a smaller asymptotic variance compared to $\widehat{\beta}_{\text{EL5}}$ because $\widehat{\beta}_{\text{EL4}}$ is based on an optimal linear combination of all the constraints whereas $\widehat{\beta}_{\text{EL5}}$ is based on the linear combination of the difference between $\mathbf{s}_\alpha(Y, \mathbf{X}, R; \alpha)$ and $R\mathbf{s}_{c,\alpha}(Y, \mathbf{X}, \mathbf{Z}; \beta, \alpha)$ and other

constraints (Newey and Smith 2004). However, the formulation in (4.11) avoids the possible collinearity and provides an alternative estimator that still has a relatively high efficiency. The exact expressions for the asymptotic variances of $\hat{\beta}_{\text{EL4}}$ and $\hat{\beta}_{\text{EL5}}$ can be found through the formula provided in Qin & Lawless (1994), and are provided in Chapter C.

4.2.3 When Zero Selection Probability for Certain Individuals is Present

We now consider two-phase studies where part of the cohort has zero probability of entering the Phase 2 sample. We are no longer able to employ the auxiliary estimating function $u(\mathbf{X}, \mathbf{Z}; \beta, \alpha, \theta)$ as Section 4.2.2, as $\pi(Y, \mathbf{X}; \alpha)$ may be zero thus cannot be in the denominator. Instead, we derive an alternative estimating function $v(\mathbf{X}, \mathbf{Z}; \beta, \alpha, \theta)$. Let \mathcal{D} denote the set of values in the range of Y that correspond to positive selection probabilities to enter Phase 2. For example, for a continuous response Y that can take any real values, suppose the two-phase ODS only samples subjects with response values smaller than a constant c_1 or larger than a constant c_2 to enter Phase 2, then $\mathcal{D} = \{y : y < c_1 \text{ or } y > c_2\}$. Let $S = I(Y \in \mathcal{D})$ denote the indicator for having a positive probability of being selected to enter Phase 2. That is, $S = 1$ if the subject has a positive probability of entering Phase 2, and $S = 0$ if the subject has a zero probability of entering Phase 2. We require \mathcal{D} to be a known region, and this is the case for most stratified two-phase studies. Then we can write $P(R_i = 1 | Y_i, \mathbf{X}_i, S_i) = S_i \pi(Y_i, \mathbf{X}_i)$, and the CML estimator maximizes the conditional likelihood corresponding to the following conditional density:

$$f_{cc}(Y | \mathbf{X}, \mathbf{Z}, S = 1, R = 1; \beta) = \frac{f(Y | \mathbf{X}, \mathbf{Z}; \beta) \pi(Y, \mathbf{X}) I(Y \in \mathcal{D})}{\int f(Y | \mathbf{X}, \mathbf{Z}; \beta) \pi(Y, \mathbf{X}) I(Y \in \mathcal{D}) dY}.$$

It is worth pointing out that, for case-only studies in the literature where Y is binary and only the cases are selected (e.g., Piegorsch et al. 1994), the conditional density $f_{cc}(Y|\mathbf{X}, \mathbf{Z}, S = 1, R = 1; \boldsymbol{\beta})$ degenerates. Therefore the CML method, and hence our proposed method, does not work for case-only studies.

To incorporate the Phase 1 information summarized as (4.2) into the estimation of $\boldsymbol{\beta}$ and have a guaranteed efficiency improvement over the CML estimator, we assume the working model $f(Y|\mathbf{X}; \boldsymbol{\theta})$ for $f(Y|\mathbf{X})$ is correctly specified so that (4.2) becomes $E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0)\} = \mathbf{0}$, where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$ such that $f(Y|\mathbf{X}; \boldsymbol{\theta}_0) = f(Y|\mathbf{X})$. Here $\boldsymbol{\theta}_0$ is still the asymptotic limit of $\hat{\boldsymbol{\theta}}$. Define

$$\begin{aligned} \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0) &= E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0)|\mathbf{X}, S = 1\} \\ &= \int \mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) \frac{P(S = 1|Y, \mathbf{X})f(Y|\mathbf{X})}{P(S = 1|\mathbf{X})} dY \\ &= \frac{\int \mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0)f(Y|\mathbf{X}; \boldsymbol{\theta}_0)I(Y \in \mathcal{D})dY}{\int f(Y|\mathbf{X}; \boldsymbol{\theta}_0)I(Y \in \mathcal{D})dY}, \end{aligned}$$

then we must have $E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)|\mathbf{X}, S = 1\} = \mathbf{0}$, which implies that

$$E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)|S = 1\} = \mathbf{0}.$$

Therefore, we have

$$\begin{aligned} \mathbf{0} &= E\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)|S = 1\} \\ &= \int P(R = 1|S = 1) \frac{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)}{\pi(Y, \mathbf{X})} \left\{ \frac{P(R = 1|Y, \mathbf{X}, \mathbf{Z}, S = 1)f(Y, \mathbf{X}, \mathbf{Z}|S = 1)}{P(R = 1|S = 1)} \right\} dY d\mathbf{X} d\mathbf{Z} \\ &= P(R = 1|S = 1)E \left\{ \frac{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)}{\pi(Y, \mathbf{X})} \middle| S = 1, R = 1 \right\} \\ &= P(R = 1|S = 1)E \left[E \left\{ \frac{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}_0) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta}_0)}{\pi(Y, \mathbf{X})} \middle| \mathbf{X}, \mathbf{Z}, S = 1, R = 1 \right\} \middle| S = 1, R = 1 \right] \\ &= P(R = 1|S = 1)E\{\mathbf{v}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)|S = 1, R = 1\}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{v}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}) = E \left\{ \frac{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta})}{\pi(Y, \mathbf{X})} \middle| \mathbf{X}, \mathbf{Z}, S = 1, R = 1 \right\} \\ \int \frac{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta}) - \mathbf{h}^*(\mathbf{X}; \boldsymbol{\theta})}{\pi(Y, \mathbf{X})} f_{cc}(Y | \mathbf{X}, \mathbf{Z}, S = 1, R = 1; \boldsymbol{\beta}) dY. \end{aligned}$$

In other words, the information in (4.2) is now summarized as

$$E_{(\mathbf{X}, \mathbf{Z} | S=1, R=1)} \{ \mathbf{v}(\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) | S = 1, R = 1 \} = \mathbf{0}, \quad (4.12)$$

where $E_{(\mathbf{X}, \mathbf{Z} | S=1, R=1)}(\cdot | S = 1, R = 1)$ is the expectation taken under the conditional covariate distribution $F(\mathbf{X}, \mathbf{Z} | S = 1, R = 1)$, which is the same as $F(\mathbf{X}, \mathbf{Z} | R = 1)$ as $R = 1$ implies $S = 1$. The moment condition (4.12) imposes a constraint on the conditional distribution $F(\mathbf{X}, \mathbf{Z} | S = 1, R = 1)$. Thus, the empirical likelihood estimators with constraints in (4.5) and (4.7) can be similarly defined as

$$\begin{aligned} \max_{\boldsymbol{\beta}, p_1, \dots, p_m} \prod_{i=1}^m f_{cc}(Y_i | \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) p_i \quad \text{subject to} \\ p_i \geq 0, \sum_{i=1}^m p_i = 1, \text{ and } \sum_{i=1}^m p_i \mathbf{v}(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^*) = \mathbf{0}, \end{aligned} \quad (4.13)$$

and

$$\max_{\boldsymbol{\beta}, p_1, \dots, p_n} \prod_{i=1}^n p_i \quad \text{subject to} \quad (4.14)$$

$$p_i \geq 0, \sum_{i=1}^m p_i = 1, \text{ and } \sum_{i=1}^m p_i \begin{pmatrix} R_i \mathbf{v}(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\theta}^*) \\ \mathbf{s}_{c, \boldsymbol{\beta}}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}) \end{pmatrix} = \mathbf{0}, \quad (4.15)$$

where $p_i = dF(\mathbf{X}_i, \mathbf{Z}_i | R_i = 1, S_i = 1)$ and $q_i = dF(Y_i, \mathbf{X}_i, \mathbf{Z}_i | R_i = 1, S_i = 1)$ respectively. Their asymptotic variances can also be similarly shown to be less than or equal to the ones

for classic CML estimator. By adding the estimating equations of α which indexes the model $\pi(Y, \mathbf{X}; \alpha) = P(R = 1|Y, \mathbf{X}, S = 1; \alpha)$, we can have similar $\hat{\beta}_{\text{EL3}}$ to $\hat{\beta}_{\text{EL5}}$ defined with empirical probabilities $q_i = dF(Y_i, \mathbf{X}_i, \mathbf{Z}_i, S_i, R_i)$ through

$$\begin{aligned} & \max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \text{ subject to} \\ & q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{cc, \beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ S_i \mathbf{s}_{\alpha}(Y_i, \mathbf{X}_i, R_i; \alpha) \\ R_i \mathbf{v}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}. \end{aligned} \quad (4.16)$$

$$\begin{aligned} & \max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \text{ subject to} \\ & q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{c, \beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{s}_{c, \alpha}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ S_i \mathbf{s}_{\alpha}(Y_i, \mathbf{X}_i, R_i; \alpha) \\ R_i \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}. \end{aligned} \quad (4.17)$$

and

$$\max_{\beta, \alpha, \theta, q_1, \dots, q_n} \prod_{i=1}^n q_i \text{ subject to}$$

$$q_i \geq 0, \sum_{i=1}^n q_i = 1, \text{ and } \sum_{i=1}^n q_i \begin{pmatrix} R_i \mathbf{s}_{cc,\beta}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ S_i \mathbf{s}_\alpha(Y_i, \mathbf{X}_i, R_i; \alpha) - R_i \mathbf{s}_{cc,\alpha}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha) \\ R_i \mathbf{v}(\mathbf{X}_i, \mathbf{Z}_i; \beta, \alpha, \theta) \\ \mathbf{h}(Y_i, \mathbf{X}_i; \theta) \end{pmatrix} = \mathbf{0}. \quad (4.18)$$

The order between the asymptotic variances of $\hat{\beta}_{\text{EL3}}$ and $\hat{\beta}_{\text{EL4}}$ and between $\hat{\beta}_{\text{EL5}}$ and $\hat{\beta}_{\text{EL4}}$ can again be directly derived from Corollary 1 of Qin & Lawless (1994). When there is no collinearity among the estimating functions, $\hat{\beta}_{\text{EL4}}$ is the most efficient, but when there is collinearity present, $\hat{\beta}_{\text{EL5}}$ and $\hat{\beta}_{\text{EL3}}$ provide highly efficient alternatives. The asymptotic variances of $\hat{\beta}_{\text{EL3}}$ to $\hat{\beta}_{\text{EL5}}$ can again be derived by directly applying the results of Qin & Lawless (1994).

4.3 Simulation Studies

Through the simulations, our EL estimators are implemented similarly to that in Chapters 2 and 3. The algorithm is outlined in Section 2.5, except that we used “fmincon” in MATLAB for the inner loop.

4.3.1 Binary Outcome in a Logistic Regression Model with Expensive Covariate

In this study, the outcome Y is a binary variable, depending on two 1-dimensional covariates X and Z , where Z is expensive to measure. We first generate (\tilde{X}_i, Z_i) $i = 1, \dots, n = 2000$ from a bivariate normal distribution, with both variables having zero mean and unit standard deviation. The correlation coefficient ρ between them is set to 0.1. X

is a categorized version of \tilde{X} , coded as 0,1,2 respectively for \tilde{X} values in three strata: $(-\infty, -0.44]$, $(-0.44, 0.44]$, $(0.44, \infty)$ which corresponds approximately to the first, second and third tertile. We generate Y_i as a realization of Bernoulli trial each with probability

$$P(Y_i = 1|X_i, Z_i) = \text{expit}(\beta_c + \beta_X X_i + \beta_Z Z_i)$$

where $\beta_0 = (-4, 1, 1)$. This results in about 10% of individuals having the outcome equal to 1.

The indicator R_i to indicate whether the i -th subject is included in the Phase 2 sample is also Bernoulli following

$$P(R_i = 1|Y_i, X_i, Z_i) = \pi_1(Y_i, X_i; \alpha) = \text{expit}(\alpha_c + \alpha_Y Y_i).$$

where α_0 is tuned to make the cases with $Y = 1$ and $Y = 0$ of approximately equal size in Phase 2. Specifically, we choose $\alpha_0 = (-3.5, 2.3)$, which results in approximately 5% individuals entering Phase 2. We consider a π model which is a logistic model with covariate Y only, and another π model to include a richer stratification of the Phase 1 sample. Specifically, we use a logistic model with covariates $(Y, I(X = 1), I(X = 2))$. We denote the two π models as “est” and “sat” and respectively name the corresponding estimators using either model.

The auxiliary $\mathbf{h}(Y, X; \theta)$ function we use here is the score function of the logistic regression model $P(Y = 1|X) = \text{expit}(\theta_c + \theta_X X)$. It is to be noted that this model is mathematically incompatible with the original model $P(Y = 1|X, Z) = \text{expit}(\beta_c + \beta_X X + \beta_Z Z)$. Table 4.1 contains the simulation results based on a Phase 1 sample size $n = 2000$ with 1000 replications.

As we can conclude that EL1 and EL2 are not as efficient as EL3 to EL5, we do not

include the performances of EL1 and EL2 in the simulations. Note that in this setup, the EL4 estimator has functionally dependent estimating functions. Similar phenomenon has been investigated in Section B.3. We therefore omit the EL4 estimator and instead compare EL3 and EL5 estimators with CML and SW. The full maximum likelihood estimator in Zhao et al. (2009) is included as the benchmark for comparison. In Table 4.1, the mean values, empirical standard errors (in round brackets) of all estimators, and the estimated standard errors (in box brackets) are given. The estimated SE's are computed from the sandwich form as in (1.12). We observe similar comparisons between CML and SW estimators as in Chapter 3. That SW has smaller SE's than CML, and a post-stratification using a richer π model improves the efficiency. For our proposed EL estimators, both the EL3 and EL5 estimator improves the efficiency of estimation of β_x and the intercept and are sometimes very close to the ML estimator. For the intercept, the SE is reduced by about 30%, and for β_x the SE's is reduced by about 40%. However, contrary to our expectation, using a more highly stratified π -est model does not help improve the efficiency. We even see an increase in the empirical SE's. For the EL3 estimators, the estimated SE's are slightly smaller. Therefore, we think that the decreases in efficiency are due to computational inaccuracy with higher dimensions of estimating functions. By using more estimating function components, there is a chance that the variations in the components are not large enough, especially when the sample size is not large enough. In the Newton method we employ to solve for EL (see 2.5), the Hessian matrix may be ill-conditioned, leading to unstable numerical performances. The unstable numerical performances may include inaccurate updating directions, or much more steps to convergence leading to prematurely stopped iterations. For the EL5-sat estimator, as the estimates may not be accurate (e.g., the estimate of the intercept has a larger bias than any other estimators), the estimated SE may not be accurate, either, which may be the probable reason that we observe a larger estimated SE. In the later Section 4.3.2 we can see this more clearly by comparing different sample sizes.

Table 4.1: Binary outcome in a logistic regression model, $n = 2000$

Method	Means (Empirical Standard Error)[Estimated Standard Error]		
	$\beta_c (\beta_{c0} = -4)$	$\beta_z (\beta_{z0} = 1)$	$\beta_x (\beta_{x0} = 1)$
CML0	-4.1265 (0.5971)	1.0515 (0.3131)	1.0603 (0.3574)
CML-est	-4.1187 (0.5473)	1.0515 (0.3131)	1.0603 (0.3574)
CML-sat	-4.1173 (0.5442)	1.0514 (0.3131)	1.0581 (0.3557)
SW-est	-4.1173 (0.5443)	1.0514 (0.3131)	1.0580 (0.3557)
SW-sat	-4.1187 (0.5473)	1.0515 (0.3131)	1.0603 (0.3574)
ML	-4.0700 (0.3688)	1.0485 (0.3158)	1.0278 (0.1977)
EL3-est	-4.0763 (0.3829)[0.3474]	1.0564 (0.3181)[0.2957]	1.0305 (0.2048)[0.1953]
EL3-sat	-4.0683 (0.3850)[0.3430]	1.0411 (0.3368)[0.2903]	1.0292 (0.2049)[0.1931]
EL5-est	-4.0763 (0.3829)[0.3476]	1.0562 (0.3181)[0.2965]	1.0305 (0.2048)[0.1953]
EL5-sat	-4.1600 (0.4031)[0.3611]	1.0323 (0.3242)[0.2952]	1.0621 (0.2153)[0.2000]

Another scenario that is of interest is when we have the inexpensive covariate X is a surrogate for the expensive covariate Z , in the sense that $Y \perp X|Z$. We include a smaller simulation with 500 replicates, each with $n = 2000$ individuals in the Appendix section, Section C.4.

4.3.2 Continuous Outcome in a Linear Regression Model with Expensive Covariate

In this study, we simulate a normal linear regression model, $f(Y|X, Z; \boldsymbol{\beta}) = N(\beta_c + \beta_X X + \beta_Z Z, \sigma^2)$, with $\boldsymbol{\beta}_0 = (0, 1, 1)$ and $\sigma = 2$. We generate Z the expensive covariate and another random variable, \tilde{X} from a bivariate normal distribution, with marginal standard normals and correlation coefficient $\rho = 0.1$. We define X as a categorization of \tilde{X} based on break-points at $(-0.44, 0.44)$, which are approximately the 1/3 and 2/3 quantiles of \tilde{X} , and coded as 0, 1, 2 in the first, second and third tertile. In phase 1, (Y, X) is observed for every individual and in Phase 2, we use VPS with $P(R = 1|Y, X) = \pi(Y; \boldsymbol{\alpha}) = \sum_{k=1}^K \alpha_k I(Y \in S_k)$, where the α_k are specified probabilities and $S_k = (c_{k-1}, c_k]$ is the k -th stratum in a partition of the

sample space for Y . We consider two Phase 2 sampling plans here. As often used in genetic epidemiological studies, we over-sample the lower and upper tails of the Y distribution. We set $K = 3$ where two cut points divide the real line into 3 strata, and the two cut points are set close to the 1st and 3rd quartiles of Y . Specifically, we choose $c_1 = -0.44$, $c_2 = 2.45$. In the first Phase 2 sampling plan, we set $\alpha_0 = (\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.05, 0.5)$, so about 33% of a Phase 1 sample is selected for Phase 2; in the second Phase 2 sampling plan, we set $\alpha_0 = (\alpha_1, \alpha_3) = (0.3, 0.5)$ (α_2 is always zero and thus omitted) so about 20% of a Phase 1 sample is selected for Phase 2.

In this scenario, the conditional density function for Y is

$$\begin{aligned} f_c(Y|X = x, Z = z; \beta, \alpha) &= \frac{\exp\{-(Y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(Y \in S_k)}{\int \exp\{-(Y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(y \in S_k) dy} \\ &= \frac{\exp\{-(Y - \beta_c - \beta_x x - \beta_z z)^2/(2\sigma^2)\} \sum_{k=1}^3 \alpha_k I(Y \in S_k)}{\sum_{k=1}^3 \alpha_k \{F(c_k - \beta_c - \beta_x x - \beta_z z) - F(c_{k-1} - \beta_c - \beta_x x - \beta_z z)\}}. \end{aligned}$$

The parameter α in the true π model can be estimated through estimating equation

$$\left(\frac{\sum_{y_i \in S_1} R_i}{\sum_i I(y_i \in S_1)} - \hat{\alpha}_1, \frac{\sum_{y_i \in S_2} R_i}{\sum_i I(y_i \in S_2)} - \hat{\alpha}_2, \frac{\sum_{y_i \in S_3} R_i}{\sum_i I(y_i \in S_3)} - \hat{\alpha}_3 \right) = \mathbf{0},$$

and

$$\left(\frac{\sum_{y_i \in S_1} R_i}{\sum_i I(y_i \in S_1)} - \hat{\alpha}_1, \frac{\sum_{y_i \in S_3} R_i}{\sum_i I(y_i \in S_3)} - \hat{\alpha}_3 \right) = \mathbf{0},$$

respectively. Post-stratification is also considered, using a more highly stratified model where we also include X in the selection model; we define 9-dimensional

$$\tilde{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}),$$

and 6-dimensional

$$\tilde{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{31}, \alpha_{32}, \alpha_{33})$$

respectively, where $\alpha_{ij} = P(R = 1|Y \in S_i, X = j - 1)$, as $\alpha_{.j}$ corresponds to $X = j - 1$ for $j = 1, 2, 3$.

We use $\mathbf{h}(Y, X; \boldsymbol{\theta}) = ((Y - \theta_c - \theta_X X), (Y - \theta_c - \theta_X X)X^T)^T$ as the auxiliary estimating function. The $\mathbf{u}(X, Z; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$, $\mathbf{h}^*(X; \boldsymbol{\theta})$ and $\mathbf{v}(X, Z; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ can be therefore defined. It can be easily verified that, when categorizing \tilde{X} into X , as long as the two cut points are symmetric about the mean of \tilde{X} , the $\mathbf{h}(Y, X; \boldsymbol{\theta})$ model is correctly specified, which is the case here.

We compare the performances of CML and SW estimators and our candidate EL estimators, EL3, EL4 and EL5, each with the true selection model as well as a richer stratification model. The semiparametric maximum likelihood estimator proposed as Zhao et al. (2009) is implemented and labeled ML as a benchmark for comparison. For $\boldsymbol{\alpha}_0 = (0.3, 0.05, 0.5)$, simulation results are shown in Table 4.2 and Table 4.3 with Phase 1 sample size $n = 2000$ and $n = 300$ respectively, and their respective m is approximately 500 and 75. For $\boldsymbol{\alpha}_0 = (0.3, 0.5)$, simulation results are shown in Table 4.4 and Table 4.5 with Phase 1 sample size $n = 2000$ and $n = 300$ respectively, and Phase 2 sample size about 400 and 60. We see our standard error estimates are very close to the empirical standard errors. When we use the non-saturated π model, namely $\pi_{est}(Y; \boldsymbol{\alpha})$, the supposed best EL estimator, E4-est further improves upon SW. EL3-est and EL5-est also have very good performances, indicating that the auxiliary function $\mathbf{u}(X, Z; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ may be a better choice than the $\mathbf{s}_{c,\alpha}(Y, X, Z; \boldsymbol{\beta}, \boldsymbol{\alpha})$. When using the saturated $\pi(Y, X; \boldsymbol{\alpha})$, SW achieves very good performance. For the same EL estimator, using saturated $\pi(Y, X; \boldsymbol{\alpha})$ slightly worsens the efficiency sometimes. The deterioration is more apparent for EL4-sat, and in this case, our variance estimates are a bit off the empirical standard errors, too. This indicates that computational issues are arising likely due to an increased number of estimating functions, similar to Section 4.3.1. The substantially higher dimensions likely lead to that the covariance between this components of estimating functions are not large enough. This is confirmed by comparing the $n = 2000$ and $n = 300$ cases.

Table 4.2: Continuous outcome in a linear regression model with positive selection probabilities, $n = 2000$

Estimator	Means (Empirical Standard Error)[Estimated Standard Error]			
	$\beta_c (\beta_{c0} = 0)$	$\beta_z (\beta_{z0} = 1)$	$\beta_x (\beta_{x0} = 1)$	$\sigma (\sigma_0 = 2)$
CML0	0.000 (0.126)	0.999 (0.083)	1.001 (0.099)	1.994 (0.056)
CML-est	-0.000 (0.118)	0.999 (0.083)	1.002 (0.099)	1.994 (0.054)
CML-sat	0.000 (0.105)	1.000 (0.082)	0.999 (0.081)	1.995 (0.054)
SW-est	-0.000 (0.114)	0.999 (0.080)	1.001 (0.095)	1.995 (0.048)
SW-sat	-0.002 (0.100)	1.003 (0.079)	1.000 (0.076)	1.997 (0.047)
ML	0.000 (0.095)	1.001 (0.080)	0.999 (0.072)	2.000 (0.043)
EL3-est	0.001 (0.095)[0.093]	1.001 (0.083)[0.082]	0.998 (0.071)[0.071]	1.996 (0.054)[0.054]
EL3-sat	0.001 (0.095)[0.093]	1.000 (0.080)[0.080]	0.998 (0.071)[0.071]	1.996 (0.050)[0.051]
EL4-est	0.001 (0.095)[0.090]	1.001 (0.080)[0.078]	0.998 (0.071)[0.070]	1.996 (0.047)[0.047]
EL4-sat	0.001 (0.096)[0.081]	1.001 (0.079)[0.076]	0.998 (0.072)[0.064]	1.996 (0.046)[0.045]
EL5-est	0.001 (0.095)[0.093]	1.001 (0.080)[0.078]	0.998 (0.071)[0.071]	1.996 (0.047)[0.047]
EL5-sat	-0.000 (0.095)[0.093]	1.001 (0.078)[0.078]	0.999 (0.071)[0.071]	1.997 (0.046)[0.046]

The efficiency decrease is more evident for a smaller sample size. Therefore, we do see a trade-off between post-stratification and less estimating equations, i.e., computational accuracy, especially for smaller sample sizes. Actually, this issue may arise for any EL estimators but is more evident for EL estimators with a large number of constraints.

For reference, we also include a simulation study for the positive selection case, by varying the correlation between covariates as well as varying the main model coefficients in Section C.5.

4.4 Illustration on a Genetics Study

We consider an illustration to genetic association testing based on data from Genetic Analysis Workshop 17 (Almasy et al. 2011) as in Section 3.4. All the working models used are the same as in Section 3.4, too. The estimates of regression coefficients in the normal regression model $N(\beta_c + \beta'_X \mathbf{X} + \beta_Z Z, \sigma)$ for Y are shown in Table 4.6. For comparison we also show the estimates based on the full data for all 321 individuals, which is available. A

Table 4.3: Continuous outcome in a linear regression model with positive selection probabilities, $n = 300$

Method	Means (Empirical Standard Error)[Estimated Standard Error]			
	$\beta_c (\beta_{c0} = 0)$	$\beta_z (\beta_{z0} = 1)$	$\beta_x (\beta_{x0} = 1)$	$\sigma (\sigma_0 = 2)$
CML0	0.027 (0.332)	1.014 (0.219)	0.974 (0.247)	1.963 (0.148)
CML-est	0.022 (0.318)	1.025 (0.218)	0.986 (0.248)	1.974 (0.148)
CML-sat	0.020 (0.276)	1.028 (0.219)	0.986 (0.206)	1.977 (0.148)
SW-est	0.015 (0.311)	1.034 (0.213)	0.993 (0.241)	1.982 (0.134)
SW-sat	0.008 (0.269)	1.042 (0.216)	0.999 (0.197)	1.994 (0.134)
ML	0.017 (0.255)	0.989 (0.204)	1.025 (0.202)	1.981 (0.111)
EL3-est	0.024 (0.247)[0.242]	1.028 (0.222)[0.206]	0.981 (0.183)[0.184]	1.978 (0.151)[0.134]
EL3-sat	0.026 (0.246)[0.241]	1.026 (0.219)[0.201]	0.980 (0.183)[0.183]	1.976 (0.142)[0.126]
EL4-est	0.025 (0.247)[0.229]	1.024 (0.214)[0.192]	0.981 (0.183)[0.180]	1.971 (0.132)[0.115]
EL4-sat	0.019 (0.250)[0.215]	1.020 (0.212)[0.183]	0.984 (0.186)[0.169]	1.968 (0.129)[0.107]
EL5-est	0.021 (0.247)[0.242]	1.032 (0.214)[0.198]	0.985 (0.183)[0.184]	1.981 (0.130)[0.120]
EL5-sat	0.014 (0.253)[0.242]	1.036 (0.213)[0.199]	0.991 (0.187)[0.184]	1.989 (0.130)[0.118]

Table 4.4: Continuous outcome in a linear regression model with zero selection probabilities, $n = 2000$

Estimator	Means (Empirical Standard Error)[Estimated Standard Error]			
	$\beta_c (\beta_{c0} = 0)$	$\beta_z (\beta_{z0} = 1)$	$\beta_x (\beta_{x0} = 1)$	$\sigma (\sigma_0 = 2)$
CML0	0.001 (0.127)	0.998 (0.087)	1.001 (0.102)	1.994 (0.058)
CML-est	0.000 (0.121)	0.999 (0.087)	1.002 (0.102)	1.994 (0.058)
CML-sat	0.001 (0.107)	1.000 (0.087)	0.999 (0.084)	1.995 (0.057)
SW-est	0.000 (0.119)	0.999 (0.087)	1.002 (0.102)	1.994 (0.058)
SW-sat	0.002 (0.105)	1.000 (0.087)	0.998 (0.082)	1.995 (0.057)
ML	0.001 (0.097)	1.002 (0.079)	0.998 (0.073)	1.998 (0.043)
EL3-est	0.002 (0.097)[0.096]	1.000 (0.088)[0.086]	0.998 (0.073)[0.073]	1.996 (0.057)[0.057]
EL3-sat	0.002 (0.097)[0.096]	1.000 (0.085)[0.085]	0.997 (0.073)[0.073]	1.996 (0.054)[0.055]
EL4-est	0.002 (0.097)[0.094]	1.000 (0.088)[0.086]	0.998 (0.073)[0.072]	1.996 (0.057)[0.058]
EL4-sat	0.001 (0.098)[0.089]	1.001 (0.085)[0.082]	0.998 (0.073)[0.068]	1.996 (0.053)[0.053]
EL5-est	0.002 (0.097)[0.096]	1.000 (0.088)[0.086]	0.997 (0.073)[0.073]	1.996 (0.057)[0.057]
EL5-sat	0.002 (0.097)[0.096]	1.001 (0.084)[0.084]	0.997 (0.073)[0.073]	1.996 (0.053)[0.054]

Table 4.5: Continuous outcome in a linear regression model with zero selection probabilities, $n = 300$

Estimator	Means (Empirical Standard Error)[Estimated Standard Error]			
	$\beta_c (\beta_{c0} = 0)$	$\beta_z (\beta_{z0} = 1)$	$\beta_x (\beta_{x0} = 1)$	$\sigma (\sigma_0 = 2)$
CML0	-0.014 (0.338)	1.028 (0.226)	1.009 (0.267)	1.967 (0.159)
CML-est	-0.010 (0.320)	1.031 (0.227)	1.012 (0.268)	1.969 (0.160)
CML-sat	0.001 (0.280)	1.037 (0.229)	0.997 (0.224)	1.974 (0.159)
SW-est	-0.010 (0.319)	1.031 (0.227)	1.012 (0.268)	1.969 (0.159)
SW-sat	-0.002 (0.274)	1.035 (0.229)	1.000 (0.213)	1.975 (0.158)
ML	0.007 (0.240)	1.030 (0.214)	0.996 (0.187)	1.977 (0.111)
EL3-est	0.007 (0.244)[0.250]	1.034 (0.233)[0.218]	0.996 (0.190)[0.190]	1.973 (0.161)[0.141]
EL3-sat	0.008 (0.244)[0.249]	1.034 (0.227)[0.214]	0.998 (0.192)[0.189]	1.973 (0.157)[0.137]
EL4-est	0.009 (0.241)[0.248]	1.032 (0.233)[0.216]	0.995 (0.188)[0.188]	1.969 (0.164)[0.138]
EL4-sat	0.005 (0.250)[0.246]	1.035 (0.229)[0.207]	0.996 (0.191)[0.188]	1.965 (0.156)[0.129]
EL5-est	0.007 (0.244)[0.251]	1.036 (0.233)[0.219]	0.996 (0.189)[0.190]	1.975 (0.161)[0.142]
EL5-sat	0.006 (0.248)[0.251]	1.038 (0.225)[0.217]	0.996 (0.191)[0.190]	1.978 (0.151)[0.137]

maximum likelihood estimate is obtained using the full data. We see that the most efficient EL estimates based on Z data for only about 30 percent of individuals have standard errors that are only moderately larger than those for the estimates from the full data. All the EL estimators, however, represent very similar standard errors, comparing to the CML and SW estimator. This may be due to that the auxiliary information are captured largely by the function $\mathbf{v}(Y, \mathbf{X}, Z; \boldsymbol{\theta})$ instead of by the functions $R\mathbf{s}_{c,\alpha}(Y, \mathbf{X}, Z; \boldsymbol{\beta}, \boldsymbol{\alpha})$ which the SW estimator employed.

4.5 Discussion

We propose a new framework based on the general empirical likelihood for general two-phase ODS, where a scalar response Y and an inexpensive covariate vector \mathbf{X} is observed for every subject, while an expensive covariate vector \mathbf{Z} is observed for only part of them. Based on the conditional likelihood, our method makes use of the Phase 1 data to improve the efficiency. It is versatile for a much wider range of two-phase designs and covariate

Table 4.6: Regression coefficient estimates for the GAW 17 two-phase data

Method	Mean [Estimated Standard Error]									
	Intercept		# of rare variants		sex		age		smoke	
CML0	-0.1541	[0.2307]	0.2773	[0.1331]	0.0442	[0.1496]	0.2717	[0.0873]	0.6390	[0.1769]
CML-est	-0.1983	[0.2305]	0.2787	[0.1350]	0.0432	[0.1500]	0.2730	[0.0886]	0.6463	[0.1806]
CML-sat	-0.1936	[0.2275]	0.2766	[0.1336]	0.0334	[0.1481]	0.3074	[0.0897]	0.6126	[0.1807]
SW-est	-0.1990	[0.2296]	0.2788	[0.1351]	0.0432	[0.1499]	0.2730	[0.0888]	0.6464	[0.1800]
SW-sat	-0.2028	[0.2296]	0.2760	[0.1332]	0.0356	[0.1500]	0.3143	[0.0815]	0.6751	[0.1613]
EL5-est	-0.2021	[0.1884]	0.3006	[0.1384]	0.0015	[0.1083]	0.3270	[0.0579]	0.6257	[0.1075]
EL5-sat	-0.2195	[0.1833]	0.3126	[0.1351]	0.0056	[0.1043]	0.3108	[0.0537]	0.6488	[0.1088]
EL6-est	-0.1989	[0.1702]	0.3027	[0.1277]	-0.0017	[0.1045]	0.3316	[0.0519]	0.6250	[0.1170]
EL6-sat	-0.2726	[0.1749]	0.2974	[0.1300]	0.0488	[0.1089]	0.2954	[0.0447]	0.6190	[0.1035]
EL7-est	-0.1969	[0.1889]	0.3047	[0.1425]	-0.0040	[0.1085]	0.3372	[0.0583]	0.6232	[0.1074]
EL7-sat	-0.2442	[0.1792]	0.3116	[0.1329]	0.0217	[0.1015]	0.3159	[0.0538]	0.6540	[0.1110]
Full data	-0.2489	[0.1701]	0.3164	[0.0857]	0.0344	[0.1011]	0.2805	[0.0505]	0.6878	[0.1176]

distributions compared to other methods. It allows a design where every subject has a positive selection probability into Phase 2, or where subjects with certain Y values will never be selected. Our methodology avoids the modelling of covariate distribution which is of practical use when we have more than one inexpensive covariates, or when the inexpensive covariates are a mix of discrete and continuous variables.

Compared to CML estimators, our new method provides a way to use Phase 1 data systematically. The auxiliary estimating function we propose is very flexible in terms of postulating any working model of $f(Y, \mathbf{X}; \boldsymbol{\theta})$. It is to be noted that such working model need not be strictly “correct” when $\pi_i > 0$ for all i . Even when $\pi_i = 0$ for some individuals, as the variables involved in this model are observed for everyone, the model can be easily checked and thus is not severely misspecified. In fact, in our simulations with logistic regression models (see Sections 4.3.1, C.4), using a working model that is mathematically incompatible with the original model of interest still yields minimal bias and significant improvement of efficiencies in many cases.

Our framework also covers existing estimators studied in Chapter 3 as special cases, where the utilization of Phase 1 data comes from the estimation of the nuisance parameter $\boldsymbol{\alpha}$ in the selection model π . In our numerical simulations, we see that the new methodology

applies to both discrete and continuous responses. However, for a continuous response, the room of improvement may be small, as classic CML with post-stratification on the selection model π is quite efficient in many settings. For binary outcomes, the improvement upon competing estimators is more significant. In many scenarios, our EL estimators are close to the efficient ML estimator. We also observe slightly worse performance sometimes when a more complex π model is employed. Specifically, we consider that there may be an issue of convergence to the global minimum. The empirical SE's are larger than the estimated ones, and some estimators theoretically more efficient shows worse efficiency instead (e.g., EL4-sat in Section 4.3.2). This declining performance suggests a trade-off between the improvement from post-stratification and numerical stability. Some alternative Newton-based optimization algorithm may be considered to implement the maximization of the profile likelihood.

We also see some possible extensions of our framework to more general settings other than two-phase samples. For example, in the data integration problems considered by Qin, Zhang, Li, Albanes & Yu (2015), Chatterjee et al. (2016) and Han & Lawless (2019), an external big data source provides a relatively accurate estimate of $\boldsymbol{\theta}$ without individual-level data. Our method may also be employed by ignoring the uncertainty in the estimation of $\boldsymbol{\theta}$, as discussed in Section 4.2.2. We look forward to developing extensions to our framework and applying it to more specific application settings.

Chapter 5

Discussions and Future Work

5.1 Discussion

5.1.1 Regression with Missing Covariates

In this thesis, we consider regression problems with a scalar outcome Y_i , and a covariate vector \mathbf{X}_i observed for every individual $i = 1, \dots, n$ of an entire cohort or representative sample. The other covariate vector \mathbf{Z}_i , on the contrary, is observed only for a subset of individuals. In general, we are interested in the estimation of a model $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ indexed by the parameter of interest $\boldsymbol{\beta}$. It can be relaxed to a more general semiparametric estimating equation, $E\{\mathbf{U}(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})\} = \mathbf{0}$ without assumptions on the specific class of distribution that $f(Y|\mathbf{X}, \mathbf{Z})$ falls in, as long as there exists a $\boldsymbol{\beta}_0$ such that $E\{\mathbf{U}(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_0)\} = \mathbf{0}$. We use an indicator variable R_i to denote if \mathbf{Z}_i is in this observed subset. For MNAR data, the ML model often has identification issues; for MAR data, ML may be difficult to postulate for \mathbf{Z} with dimension higher than 1, or when \mathbf{X} has continuous components; meanwhile, the AIPW estimator is not applicable when $\pi_i = 0$ for certain individual i (see our discussion in Section 4.1). Therefore, alternative approaches to efficiently estimate the parameter of

interest are desired.

5.1.2 Empirical Likelihood Frameworks for Exploiting Auxiliary Information

Based on the specific missing mechanism, we can write out an estimating equation of the parameter of interest β using all the complete cases, namely

$$\sum_{i=1}^n R_i U_{CC}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, R_i; \beta) = \mathbf{0}.$$

In Chapter 2, we assumed the missingness subject to $R \perp Y | \mathbf{X}, \mathbf{Z}$, which makes makes this approach valid without any transformation or weighting. In Chapters 3-4, we assume the missingness satisfies $R \perp \mathbf{Z} | Y, \mathbf{X}$, and this enables the usage of the classic CML estimator, with U_{CC} equal to the conditional score function $\mathbf{s}_{c,\beta}$. There may be some other CC based estimators, such as IPW, but in our specific scenarios IPW is found to be less efficient (Bartlett et al. 2014; Scott & Wild 2011) thus is not considered in detail.

The CC based estimators, though consistent under the corresponding missingness assumptions, are not efficient, as they are not using any information contained in the partially observed cases. In our estimation frameworks, an auxiliary estimating function $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ is used to enhance the original estimator defined through U_{CC} . It may involve both the parameter of interest β and a nuisance parameter θ . Depending on the model assumptions, the choices of $\mathbf{h}(Y, \mathbf{X}, R; \beta, \theta)$ may be very flexible. A natural choice is to fit a regression model $E(Y | \mathbf{X}; \theta) = \mu(\mathbf{X}, \theta)$. Depending on the specific assumption, we may need different assumptions to guarantee the consistency for the estimation of β . For example, in Chapter 4, we may need a correct $f(Y | \mathbf{X}; \theta)$ when some Phase 1 individuals has zero probability of entering Phase 2, but when the selection probability is always positive, an estimating equa-

tion with an asymptotic limit $\boldsymbol{\theta}^*$ is sufficient. Some other estimating functions concerning the selection model $\pi(Y, \mathbf{X}; \boldsymbol{\alpha})$ are also applicable. We found that an augmented system of estimating equations

$$\mathbf{g}(Y, \mathbf{X}, \mathbf{Z}, R; \boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} RU_{CC}(Y, \mathbf{X}, \mathbf{Z}, R; \boldsymbol{\beta}, \boldsymbol{\theta}) \\ \mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta}) \end{pmatrix}$$

improves the estimation efficiency when $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ effectively brings the information contained in the partially observed cases into the estimation. Moreover, even when the working model is slightly misspecified, or incompatible with the model of interest, we have numerical evidence to show that the EL framework has some robustness (e.g., Sections 2.5, 4.3.1).

A key point in our EL-based frameworks is to create an “over-identified” system, with the number of equations greater than the number of parameters. Otherwise, the EL procedure will end up with trivial empirical probabilities, $\hat{p}_i = 1/n$, and the estimate for $\boldsymbol{\beta}$ is identical to $\hat{\boldsymbol{\beta}}_{CC}$. It can also be interpreted as that when the added dimension of estimating functions is equal to the dimension of extra parameters, the information is all used to estimate the nuisance parameter and will not contribute to the estimation of the parameter of interest. For special cases such as the MNAR covariates satisfying the “outcome independent” assumption as in Chapter 2, we can develop an optimal form of such auxiliary estimating functions.

For two-phase ODS, directly using the regression model $E(Y|\mathbf{X}; \boldsymbol{\theta}) = \mu(Y, \mathbf{X}; \boldsymbol{\theta})$ may also be a straightforward choice. However, as the corresponding estimating function does not involve $\boldsymbol{\beta}$, the parameter of interest, this naive estimating function might not give a significant efficiency improvement. On the other hand, the parametric assumption for $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ enables us to evaluate the conditional expectation of $\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta})$ given \mathbf{X}, \mathbf{Z} , thus gives a more consolidated version of auxiliary estimating function, $\mathbf{u}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}, \boldsymbol{\theta}) =$

$E_{(\mathbf{X}, \mathbf{Z})}[E_{(Y|\mathbf{X}, \mathbf{Z})}\{\mathbf{h}(Y, \mathbf{X}; \boldsymbol{\theta})|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}\}]$. Evidences in the literature show that using such a transformed $\mathbf{u}(Y, \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}, \boldsymbol{\theta})$ gives desirable improvement for regression problems in a similar data integration problem (Qin et al. 2015; Chatterjee et al. 2016; Han & Lawless 2019).

5.1.3 Numerical Implementation of Empirical Likelihood for General Estimating Equations

Numerical implementations of EL estimations are usually based on Newton-Raphson methods; see, for example, Hall & La Scala (1990). However, due to the underlying constrained maximization being nonconvex, reliable numerical methods could be difficult to find. A commonly recommended algorithm is the saddle-point method as in Owen (2001), Han & Lawless (2019), among others. Our numerical implementations are largely based on this method, as described in Section 2.5 and 3.3, etc. We employ an inner loop to compute the Lagrange multiplier $\boldsymbol{\lambda}$ which maximized the empirical likelihood given a fixed parameter $\boldsymbol{\phi}$; an outer loop to update the parameter $\boldsymbol{\phi}$ given a fixed Lagrange multiplier $\hat{\boldsymbol{\lambda}}$. The inner one could employ many constrained optimization packages readily available, such as *constrOptim* in R, or *fmincon* in MATLAB. The outer loop can be implemented through Newton-Raphson.

A common issue arising in the problems we discussed in this thesis is the collinearity or linear dependence among different components of the estimating equations. An analytical example is given in Section B.3 of Chapter 3, where we see that in the 8 dimensions of estimating functions, the effective dimension is only 5. This issue is also commonly encountered for other settings with binary outcomes (e.g. Section 4.3.1). In these cases, directly using the Newton-Raphson method for the outer loop is problematic as the approximated Hessian matrix we need to invert is the covariance matrix of the estimating functions, which is not invertible.

Even with estimating functions linearly independent, when the number of constraints is large, and the sample size not large enough, we may not have enough variability in the estimating functions, and thus have an ill-conditioned Hessian matrix, leading to unstable numerical performances. Specifically, the Hessian matrix involves terms such as $\sum_{i=1}^n \mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\phi}) \mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\phi})^T$, and when the variation of $\mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\phi})$ is not large enough, the Newton method may run into large errors when computing the direction of updates. Some alternative computational techniques may be used to alleviate this issue, such as stochastic gradient-descent, or batch gradient-descent (Zhang 2004), but they may also bring other computational issues such as more steps to convergence. Another consideration is actually the trade-off between theoretical efficiency and numerical accuracy. This is especially the case when we use post-stratification in two-phase samples. Mathematically, a finer stratification always leads to higher efficiencies, however, on the other hand, it means more nuisance parameters to be estimated. If these parameters are to be estimated in EL, one needs to be careful in choosing a proper degree of stratification. Naively increasing the dimensions of the auxiliary estimating equations may be detrimental to the numerical performance.

5.1.4 Model Checking and Model Compatibility

As mentioned above, one advantage of auxiliary estimating function $\mathbf{h}(Y, \mathbf{X}, R; \boldsymbol{\beta}, \boldsymbol{\theta})$ is that it uses only the fully observed variables, and thus the model can be checked. In Chapter 2 we discussed the issue of model misspecification and showed an example of model checking in Section 2.5. This routine consisting of observations, P-P plots and likelihood ratio tests may be used as a general one when identifying proper working models. More specific model diagnosis for empirical likelihood for GEEs has also been proposed (Zhu, Ibrahim, Tang & Zhang 2008) which we may consider to employ.

Though diagnosis can be done, we may still have the issue of model compatibility. Natural choices of working models may not be mathematically compatible with the original model of interest. A straightforward example is that when we have a binary outcome Y , and the model of interest is a logistic regression with covariates \mathbf{Z} and \mathbf{X} , then a reduced logistic regression model of Y given \mathbf{X} is mathematically incompatible with the original model of interest. While many of our theoretical results rely on the correct specification of the working model, our simulations show that a mildly incompatible model rarely causes any bias or decrease in efficiency. Given the fact that no model is completely correct, we believe our proposed estimators have some robustness to mild model misspecifications. However, when the estimating function is unbounded, the EL estimator may not be \sqrt{n} -consistent. More robust tools are desired. For example, Schennach et al. (2007) proposed the so-called exponentially tilted EL estimator which preserves the \sqrt{n} -consistency as well as the same second order properties. This can be even further refined asymptotically by a bootstrap-based inference framework (Lee 2016).

5.2 Future Work

Checking Model Assumptions Our model assumptions include distributional assumptions such as the model of $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ or assumptions to the missingness mechanism. Checking these assumptions is an important and interesting question. In Chapter 2, we do not postulate distributional assumptions but do require conditional independence of Y and R given \mathbf{Z} and \mathbf{X} . In our application example, R represents nonresponse in alcohol consumption Z , Y is the blood pressure. As Y is measured after the collection of Z , it is reasonable to assume such conditional independence. However, in some other cases, it might need more rigorous checking for this outcome independence property to allow for this CC-based framework. In Chapters 3 and 4, for two-phase ODS, as the phase 2 selection is by

design, the missingness mechanism usually does not need to be checked. However the distributional assumption of $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ such as normal distribution may need to be checked. Sensitivity analysis is often suggested to test for a missingness mechanism assumption, but for some specific missingness types such as what we assumed in Chapter 2, developing a specific checking framework may be desired. There have been a number of studies done in sensitivity analysis to check the assumption of MAR vs MNAR (Enders 2011; Hsu, He, Hu & Zhou 2020). Many of them is based on the fact that the joint likelihood of MAR data can be written either in a selection model or in a pattern mixture model form (see our introduction in Section 1.2). We may aim to develop similar theories under the outcome-independence assumption in Chapter 2.

Robustness against model misspecifications As discussed in the section above, our methods showed good robustness against mild model misspecifications and model incompatibilities. For MLE estimators, the misspecification question has been studied as early as White (1982). For EL, we see only sparse studies on this topic, for example, Zhu et al. (2008). The behaviour of the EL estimators under model misspecification is not well studied. More investigations into the performance of EL estimators in presence of model misspecifications and incompatibilities are needed.

Comparison between the generalized method of moments and empirical likelihood

As discussed in Section 5.1.3, the numerical implementation of EL may suffer from unstable performances, especially for large numbers of estimating functions and a relatively small sample size. Besides using modified Newton methods such as stochastic gradient descent, another alternative is to use the generalized method of moments (GMM, Hansen, 1982). GMM solves an over-identified system using an iterative algorithm, and the implementation is usually less complicated than EL. In Newey & Smith (2004) the authors showed that, EL

has more favourable higher-order asymptotic properties over the GMM method. Specifically, the asymptotic bias of GMM often grows linearly with the number of over-identifying restrictions; and after bias correction, the empirical probabilities obtained from EL is higher-order efficient relative to other biased corrected estimators. However, as noted by (Imbens 2002), for a sample of size n , and a p -dimensional parameter θ to be estimated, the two-step GMM is an optimization on a p -dimensional space; whereas EL is an constrained optimization in a $(p + n)$ -dimensional space plus $r + 1$ constraints (r is the number of estimating equations). Obviously, the computation of EL is more demanding. In our practice, we also observe a trade-off between the estimation efficiency and the numerical accuracy; see, for example, Section 4.3.2. Examining the GMM method and comparing it with the EL approach in our frameworks would be of interest. We are interested in providing a general recommendation to help decide which of GMM and EL would be a better suitable choice under different circumstances.

Integrating external, summary-level information into an internal study The first extension direction is for the two-phase ODS. We see that when we ignore the uncertainty of $\hat{\theta}$ for the Phase 1 data, the EL0-1 estimator does not use any individual-level information from Phase 1, but only an estimate of the nuisance parameter. This naturally extends to data integration problems where one has access to an external big data source, but with only summary-level information. This setting is practical when the Phase 1 sample is much larger than phase 2. Further, suppose we have both a two-phase sample and an external big data source with summary level data, for example, as in Qin et al. (2015), Chatterjee et al. (2016) and Han & Lawless (2019). The Phase 1 data can be used to test for the heterogeneity in the covariate distributions of the internal two-phase sample and the external data source.

Extension to other types of regressions The second direction is to extend the current mean regression to more regression settings, such as quantile regression (QR) or functional regression problems.. For example, Tang & Leng (2012) first considered using empirical

likelihood to improve the efficiency of QR estimators when auxiliary information is available. Their method is a two-step estimator where EL is only applied to the auxiliary estimating equations. If we can assume QR models on different quantile levels, it may be still possible to get an approximated parametric model $\hat{f}(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$ (Wei, Ma & Carroll 2012) and apply our transformed estimating function \mathbf{u} . However, the computation involved may be much more expensive than the mean model case, where we have an explicit form for $f(Y|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta})$. Moreover, the estimating function of QR is not a smooth function (Koenker & Hallock 2001), which may bring in even more computational issues.

MNAR scenarios to causal inference Causal inference overlaps with missing data by treating the counterfactual outcome as missing. However, most causal inference framework requires the assumption that there are no unmeasured confounding effects. It is a comparable assumption as the MAR. Recently, there have been studies noticing that if the missing confounding effect is independent of the outcome, which is similar to the assumption we made in Chapter 2, then the average treatment effect (ATE) can be estimated without identifiability issues (Yang, Wang and Ding 2019). Yang et al. focused on the identification problem in causal inference, however, our results in Chapter 2 may help improve the estimation efficiency of the ATE.

The above are some possible extension works. We look forward to making more methodological developments based on the existing works and applying them to more specific scientific problems.

References

- Allison, P. D. (2000), “Multiple imputation for missing data: A cautionary tale,” *Sociological Methods & Research*, 28(3), 301–309.
- Anderson, T. W. (1957), “Maximum likelihood estimates for a multivariate normal distribution when some observations are missing,” *Journal of the American Statistical Association*, 52(278), 200–203.
- Barnett, I. J., Lee, S., & Lin, X. (2013), “Detecting rare variant effects using extreme phenotype sampling in sequencing association studies,” *Genetic Epidemiology*, 37(2), 142–151.
- Bartlett, J. W., Carpenter, J. R., Tilling, K., & Vansteelandt, S. (2014), “Improving upon the efficiency of complete case analysis when covariates are MNAR,” *Biostatistics*, 15(4), 719–730.
- Bjørnland, T., Bye, A., Ryeng, E., Wisløff, U., & Langaas, M. (2018), “Powerful extreme phenotype sampling designs and score tests for genetic association studies,” *Statistics in Medicine*, 37(28), 4234–4251.
- Borgan, Ø., & Samuelsen, S. O. (2013), “Nested case-control and case-cohort studies,” *Handbook of Survival Analysis*, pp. 343–367.

- Breslow, N., Day, N. et al. (1980), “Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.,” *International Agency for Research on Cancer Scientific Publications*, 1(32).
- Breslow, N. E. (1996), “Statistics in epidemiology: the case-control study,” *Journal of the American Statistical Association*, 91(433), 14–28.
- Breslow, N. E., & Holubkov, R. (1997*a*), “Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), 447–461.
- Breslow, N. E., & Holubkov, R. (1997*b*), “Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data,” *Statistics in Medicine*, 16(1), 103–116.
- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., & Kulich, M. (2009), “Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology,” *Statistics in Biosciences*, 1(1), 32–49.
- Breslow, N., Zhao, L., Fears, T. R., & Brown, C. C. (1988), “Logistic regression for stratified case-control studies,” *Biometrics*, 44(3), 891–899.
- Chan, K. C. G. (2013), “A simple multiply robust estimator for missing response problem,” *Stat*, 2(1), 143–149.
- Chatterjee, N., Chen, Y.-H., & Breslow, N. E. (2003), “A pseudoscore estimator for regression problems with two-phase sampling,” *Journal of the American Statistical Association*, 98(461), 158–168.

- Chatterjee, N., Chen, Y.-H., Maas, P., & Carroll, R. J. (2016), “Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources,” *Journal of the American Statistical Association*, 111(513), 107–117.
- Che, M., Han, P., & Lawless, J. F. (2020), “Improving estimation efficiency for regression with MNAR covariates,” *Biometrics*, 76(1), 270–280.
- Che, M., Lawless, J. F., & Han, P. (2020), “Empirical and conditional likelihoods for two-phase studies,” *Canadian Journal of Statistics*, .
- Chen, H. Y., & Li, M. (2011), “Improving power and robustness for detecting genetic association with extreme-value sampling design,” *Genetic Epidemiology*, 35(8), 823–830.
- Chen, Y.-H., & Chen, H. (2000), “A unified approach to regression analysis under double-sampling designs,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3), 449–460.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38.
- Derkach, A., Lawless, J. F., Sun, L. et al. (2014), “Pooled association tests for rare genetic variants: a review and some new results,” *Statistical Science*, 29(2), 302–321.
- DiCiccio, T., Hall, P., & Romano, J. (1991), “Empirical likelihood is Bartlett-correctable,” *the Annals of Statistics*, pp. 1053–1061.
- Enders, C. K. (2011), “Missing not at random models for latent growth curve analyses,” *Psychological Methods*, 16(1), 1.
- Hall, P., & La Scala, B. (1990), “Methodology and algorithms of empirical likelihood,” *International Statistical Review/Revue Internationale de Statistique*, pp. 109–127.

- Han, P. (2014), “Multiply robust estimation in regression analysis with missing data,” *Journal of the American Statistical Association*, 109(507), 1159–1173.
- Han, P. (2016), “Combining inverse probability weighting and multiple imputation to improve robustness of estimation,” *Scandinavian Journal of Statistics*, 43(1), 246–260.
- Han, P. (2018), “Calibration and multiple robustness when data are missing not at random,” *Statistica Sinica*, 28(4), 1725–1740.
- Han, P., & Lawless, J. F. (2016), “Discussion of constrained maximum likelihood estimation for model calibration using summary-level information from external big data source by Chatterjee, Chen, Maas and Carroll,” *Journal of the American Statistical Association*, 111, 118–121.
- Han, P., & Lawless, J. F. (2019), “Empirical likelihood estimation using auxiliary summary information with different covariate distributions,” *Statist. Sinica*, .
- Han, P., & Wang, L. (2013), “Estimation with missing data: beyond double robustness,” *Biometrika*, 100(2), 417–430.
- Hansen, L. P. (1982), “Large sample properties of generalized method of moments estimators,” *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.
- Heitjan, D. F., & Rubin, D. B. (1991), “Ignorability and coarse data,” *The Annals of Statistics*, pp. 2244–2253.
- Holcroft, C. A., Rotnitzky, A., & Robins, J. M. (1997), “Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates,” *Journal of Statistical Planning and Inference*, 65(2), 349–374.

- Holman, R., & Glas, C. A. (2005), “Modelling non-ignorable missing-data mechanisms with item response theory models,” *British Journal of Mathematical and Statistical Psychology*, 58(1), 1–17.
- Horton, N. J., & Laird, N. M. (1999), “Maximum likelihood analysis of generalized linear models with missing covariates,” *Statistical Methods in Medical Research*, 8(1), 37–50.
- Hsu, C.-H., He, Y., Hu, C., & Zhou, W. (2020), “A multiple imputation-based sensitivity analysis approach for data subject to missing not at random,” *Statistics in Medicine*, 39(26), 3756–3771.
- Hu, J., Lawless, J. F. et al. (1997), “Pseudolikelihood estimation in a class of problems with response-related missing covariates,” *Canadian Journal of Statistics*, 25(2), 125–142.
- Huang, B., & Lin, D. (2007), “Efficient association mapping of quantitative trait loci with selective genotyping,” *The American Journal of Human Genetics*, 80(3), 567–576.
- Huang, C.-Y., Qin, J., & Tsai, H.-T. (2016), “Efficient estimation of the Cox model with auxiliary subgroup survival information,” *Journal of the American Statistical Association*, 111(514), 787–799.
- Ibrahim, J. G., Lipsitz, S. R., & Chen, M.-H. (1999), “Missing covariates in generalized linear models when the missing data mechanism is non-ignorable,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1), 173–190.
- Imbens, G. W. (2002), “Generalized method of moments and empirical likelihood,” *Journal of Business & Economic Statistics*, 20(4), 493–506.
- Imbens, G. W., & Lancaster, T. (1994), “Combining micro and macro data in microeconomic models,” *The Review of Economic Studies*, 61(4), 655–680.

- Kang, J. D., Schafer, J. L. et al. (2007), “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data,” *Statistical Science*, 22(4), 523–539.
- Keogh, R. H., & Cox, D. R. (2014), *Case-control studies*, Vol. 4, Cambridge, UK: Cambridge University Press.
- Koenker, R., & Hallock, K. F. (2001), “Quantile regression,” *Journal of Economic Perspectives*, 15(4), 143–156.
- Lawless, J. (1997), “Likelihood and pseudo likelihood estimation based on response-biased observation,” *Lecture Notes-Monograph Series*, pp. 43–55.
- Lawless, J., Kalbfleisch, J., & Wild, C. (1999), “Semiparametric methods for response-selective and missing data problems in regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2), 413–438.
- Lee, S. (2016), “Asymptotic refinements of a misspecification-robust bootstrap for GEL estimators,” *Journal of Econometrics*, 192(1), 86–104.
- Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., & Conti, D. (2011), “Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies,” *Genetic Epidemiology*, 35(8), 790–799.
- Liang, H., Wang, S., & Carroll, R. J. (2007), “Partially linear models with missing response variables and error-prone covariates,” *Biometrika*, 94(1), 185–198.
- Lin, D., Hu, Y., & Huang, B. (2008), “Simple and efficient analysis of disease association with missing genotype data,” *The American Journal of Human Genetics*, 82(2), 444–452.

- Lin, D.-Y., Zeng, D., & Tang, Z.-Z. (2013), “Quantitative trait analysis in sequencing studies under trait-dependent sampling,” *Proceedings of the National Academy of Sciences*, 110(30), 12247–12252.
- Little, R. J., & Rubin, D. B. (2014), *Statistical analysis with missing data*, Vol. 333, Hoboken, New Jersey, USA: John Wiley & Sons.
- Lumley, T., Shaw, P. A., & Dai, J. Y. (2011), “Connections between survey calibration estimators and semiparametric models for incomplete data,” *International Statistical Review*, 79(2), 200–220.
- Miao, W., Ding, P., & Geng, Z. (2016), “Identifiability of normal and normal mixture models with nonignorable missing data,” *Journal of the American Statistical Association*, 111(516), 1673–1683.
- Molenberghs, G., & Kenward, M. (2007), *Missing data in clinical studies*, Vol. 61, West Sussex, England: John Wiley & Sons.
- Nelder, J., & Lee, Y. (1992), “Likelihood, quasi-likelihood and pseudolikelihood: some comparisons,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 273–284.
- Newey, W. K., & Smith, R. J. (2004), “Higher order properties of GMM and generalized empirical likelihood estimators,” *Econometrica*, 72(1), 219–255.
- Neyman, J. (1938), “Contribution to the theory of sampling human populations,” *Journal of the American Statistical Association*, 33(201), 101–116.
- Owen, A. (1990), “Empirical likelihood ratio confidence regions,” *The Annals of Statistics*, pp. 90–120.

- Owen, A. B. (1988), “Empirical likelihood ratio confidence intervals for a single functional,” *Biometrika*, 75(2), 237–249.
- Owen, A. B. (2001), *Empirical likelihood*, Boca Raton, FL, USA: Chapman and Hall/CRC.
- Padmanabhan, S., Melander, O., Johnson, T., Di Blasio, A. M., Lee, W. K., Gentilini, D., Hastie, C. E., Menni, C., Monti, M. C., Delles, C. et al. (2010), “Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension,” *PLoS Genet*, 6(10), e1001177.
- Pepe, M. S., & Fleming, T. R. (1991), “A nonparametric method for dealing with mismeasured covariate data,” *Journal of the American Statistical Association*, 86(413), 108–113.
- Qin, J. (2000), “Combining parametric and empirical likelihoods,” *Biometrika*, 87(2), 484–490.
- Qin, J. (2017), *Biased sampling, over-identified parameter problems and beyond*, Singapore: Springer Nature.
- Qin, J., & Lawless, J. (1994), “Empirical likelihood and general estimating equations,” *The Annals of Statistics*, pp. 300–325.
- Qin, J., Zhang, B., & Leung, D. H. (2009), “Empirical likelihood in missing data problems,” *Journal of the American Statistical Association*, 104(488), 1492–1503.
- Qin, J., Zhang, H., Li, P., Albanes, D., & Yu, K. (2015), “Using covariate-specific disease prevalence information to increase the power of case-control studies,” *Biometrika*, 102(1), 169–180.
- Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003), “Multiple imputation for statistical disclosure limitation,” *Journal of Official Statistics*, 19(1), 1.

- Rivera-Rodriguez, C., Haneuse, S., Wang, M., & Spiegelman, D. (2020), “Augmented pseudo-likelihood estimation for two-phase studies,” *Statistical Methods in Medical Research*, 29(2), 344–358.
- Robins, J. M., & Rotnitzky, A. (1995), “Semiparametric efficiency in multivariate regression models with missing data,” *Journal of the American Statistical Association*, 90(429), 122–129.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994), “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American statistical Association*, 89(427), 846–866.
- Rosenbaum, P. R., & Rubin, D. B. (1983), “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70(1), 41–55.
- Rotnitzky, A., & Robins, J. (1997), “Analysis of semi-parametric regression models with non-ignorable non-response,” *Statistics in Medicine*, 16(1), 81–102.
- Rotnitzky, A., & Robins, J. M. (2005), “Inverse probability weighting in survival analysis,” *Encyclopedia of Biostatistics*, .
- Rubin, D. B. (1976), “Inference and missing data,” *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1996), “Multiple imputation after 18+ years,” *Journal of the American statistical Association*, 91(434), 473–489.
- Schafer, J. L., & Graham, J. W. (2002), “Missing data: our view of the state of the art.,” *Psychological methods*, 7(2), 147.
- Schennach, S. M. et al. (2007), “Point estimation with exponentially tilted empirical likelihood,” *The Annals of Statistics*, 35(2), 634–672.

- Scott, A. J., & Wild, C. J. (1997), “Fitting regression models to case-control data by maximum likelihood,” *Biometrika*, 84(1), 57–71.
- Scott, A. J., & Wild, C. J. (2011), “Fitting regression models with response-biased samples,” *Canadian Journal of Statistics*, 39(3), 519–536.
- Seaman, S. R., & Vansteelandt, S. (2018), “Introduction to double robust methods for incomplete data,” *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2), 184.
- Seaman, S. R., & White, I. R. (2013), “Review of inverse probability weighting for dealing with missing data,” *Statistical Methods in Medical Research*, 22(3), 278–295.
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009), “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls,” *BMJ*, 338, b2393.
- Tang, C. Y., & Leng, C. (2012), “An empirical likelihood approach to quantile regression with auxiliary information,” *Statistics & Probability Letters*, 82(1), 29–36.
- Tang, G., Little, R. J., & Raghunathan, T. E. (2003), “Analysis of multivariate missing data with nonignorable nonresponse,” *Biometrika*, 90(4), 747–764.
- Tang, N., Zhao, P., & Zhu, H. (2014), “Empirical likelihood for estimating equations with nonignorable missing data,” *Statistica Sinica*, 24(2), 723.
- Tao, R., Zeng, D., & Lin, D.-Y. (2017), “Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies,” *Journal of the American Statistical Association*, 112(520), 1468–1476.
- Tao, R., Zeng, D., & Lin, D.-Y. (2019), “Optimal designs of two-phase studies,” *Journal of the American Statistical Association*, pp. 1–14.

- Tsiatis, A. (2007), *Semiparametric theory and missing data*, New York, NY, USA: Springer Science & Business Media.
- Wallace, C., Chapman, J. M., & Clayton, D. G. (2006), “Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping,” *The American Journal of Human Genetics*, 78(3), 498–504.
- Wang, D., & Chen, S. X. (2009), “Empirical likelihood for estimating equations with missing values,” *The Annals of Statistics*, pp. 490–517.
- Wang, K., Edmondson, A. C., Li, M., Gao, F., Qasim, A. N., Devaney, J. M., Burnett, M. S., Waterworth, D. M., Mooser, V., Grant, S. F. et al. (2011), “Pathway-wide association study implicates multiple sterol transport and metabolism genes in HDL cholesterol regulation,” *Frontiers in Genetics*, 2, 41.
- Wang, Q., & Rao, J. (2002), “Empirical likelihood-based inference under imputation for missing response data,” *Annals of Statistics*, pp. 896–924.
- Weaver, M. A., & Zhou, H. (2005), “An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling,” *Journal of the American Statistical Association*, 100(470), 459–469.
- Wei, Y., Ma, Y., & Carroll, R. J. (2012), “Multiple imputation in quantile regression,” *Biometrika*, 99(2), 423–438.
- White, H. (1982), “Maximum likelihood estimation of misspecified models,” *Econometrica: Journal of the Econometric Society*, pp. 1–25.
- Wu, C., & Sitter, R. R. (2001), “A model-calibration approach to using complete auxiliary information from survey data,” *Journal of the American Statistical Association*, 96(453), 185–193.

- Xie, Y., & Zhang, B. (2017), “Empirical Likelihood in Nonignorable Covariate-Missing Data Problems,” *The international journal of biostatistics*, 13(1).
- Yang, S., Wang, L., & Ding, P. (2019), “Causal inference with confounders missing not at random,” *Biometrika*, 106(4), 875–888.
- Yilmaz, Y. E., & Bull, S. B. (2011), “Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants?,” *BMC proceedings*, 5(S9), S111.
- Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988), “Models for longitudinal data: a generalized estimating equation approach,” *Biometrics*, pp. 1049–1060.
- Zeng, D., & Lin, D. (2014), “Efficient estimation of semiparametric transformation models for two-phase cohort studies,” *Journal of the American Statistical Association*, 109(505), 371–383.
- Zhang, T. (2004), Solving large scale linear prediction problems using stochastic gradient descent algorithms,, in *Proceedings of the twenty-first international conference on Machine learning*, p. 116.
- Zhang, Z., & Rockette, H. E. (2005), “On maximum likelihood estimation in parametric regression with missing covariates,” *Journal of Statistical Planning and Inference*, 134(1), 206–223.
- Zhang, Z., & Rockette, H. E. (2006), “Semiparametric maximum likelihood for missing covariates in parametric regression,” *Annals of the Institute of Statistical Mathematics*, 58(4), 687–706.
- Zhao, Y., Lawless, J. F., & McLeish, D. L. (2009), “Likelihood methods for regression models with expensive variables missing by design,” *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1), 123–136.

Zhu, H., Ibrahim, J. G., Tang, N., & Zhang, H. (2008), “Diagnostic measures for empirical likelihood of general estimating equations,” *Biometrika*, 95(2), 489–507.

APPENDICES

In this part, we report supplementary materials associated with Chapters 2-4, including proofs, mathematical derivations and additional numerical results.

Appendix A

Detailed Proofs of Theorems in Chapter 2

A.1 Proof of Theorem 1

Proof. The estimating equation is

$$\sum_{i=1}^n \begin{pmatrix} R_i \mathbf{U}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) \\ \mathbf{h}(Y_i, \mathbf{X}_i, R_i; \boldsymbol{\beta}, \boldsymbol{\theta}) \end{pmatrix} = \mathbf{0},$$

so by Qin and Lawless (1994), the asymptotic covariance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ can be written as

$$\begin{aligned} \text{ACov} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\theta}} \end{pmatrix} &= \left[\begin{pmatrix} ER\mathbf{U}_{\boldsymbol{\beta}} & E\mathbf{h}_{\boldsymbol{\beta}} \\ \mathbf{0} & E\mathbf{h}_{\boldsymbol{\theta}} \end{pmatrix}^T \begin{pmatrix} ER\mathbf{U}\mathbf{U}^T & ER\mathbf{U}\mathbf{h}^T \\ E\mathbf{h}\mathbf{U}^T & E\mathbf{h}\mathbf{h}^T \end{pmatrix}^{-1} \begin{pmatrix} ER\mathbf{U}_{\boldsymbol{\beta}} & \mathbf{0} \\ E\mathbf{h}_{\boldsymbol{\beta}} & E\mathbf{h}_{\boldsymbol{\theta}} \end{pmatrix} \right]^{-1} \\ &= \left[\begin{pmatrix} ER\mathbf{U}_{\boldsymbol{\beta}} & E\mathbf{h}_{\boldsymbol{\beta}} \\ \mathbf{0} & E\mathbf{h}_{\boldsymbol{\theta}} \end{pmatrix}^T \begin{pmatrix} U^{11} & U^{12} \\ U^{21} & U^{22} \end{pmatrix} \begin{pmatrix} ER\mathbf{U}_{\boldsymbol{\beta}} & \mathbf{0} \\ E\mathbf{h}_{\boldsymbol{\beta}} & E\mathbf{h}_{\boldsymbol{\theta}} \end{pmatrix} \right]^{-1} \end{aligned}$$

$$= \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}^{-1},$$

where

$$U^{11} = (ERUU^T)^{-1} + (ERUU^T)^{-1}ERU\mathbf{h}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} \\ ER\mathbf{h}U^T(ERUU^T)^{-1} \quad (\text{A.1})$$

$$U^{12} = (ERUU^T)^{-1}ERU\mathbf{h}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1}$$

$$U^{21} = U^{12T}$$

$$U^{22} = \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1}$$

so then

$$\begin{aligned} I_{11} &= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} \\ &+ ERU_{\beta}^T(ERUU^T)^{-1}ERU\mathbf{h}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} \\ &\cdot ER\mathbf{h}U^T(ERUU^T)^{-1}ERU_{\beta} \\ &+ E\mathbf{h}_{\beta}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} ER\mathbf{h}U^T(ERUU^T)^{-1}ERU_{\beta} \\ &+ ERU_{\beta}^T(ERUU^T)^{-1}ERU\mathbf{h}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} E\mathbf{h}_{\beta} \\ &+ E\mathbf{h}_{\beta}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} E\mathbf{h}_{\beta} \\ &= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} \\ &+ (ERU_{\beta}^T(ERUU^T)^{-1}ERU\mathbf{h}^T + E\mathbf{h}_{\beta}^T) \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} \\ &\cdot (ER\mathbf{h}U^T(ERUU^T)^{-1}ERU_{\beta} + E\mathbf{h}_{\beta}) \\ &:= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} + D_1C^{-1}D_1^T \end{aligned}$$

where we denote

$$D_1 := ERU_{\beta}^T(ERUU^T)^{-1}ERU\mathbf{h}^T + E\mathbf{h}_{\beta}^T,$$

and

$$C := E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T.$$

Meanwhile, (A.2)

$$\begin{aligned} I_{12} &= ERU_{\beta}^T(ERUU^T)^{-1}ERU\mathbf{h}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} E\mathbf{h}_{\theta} \\ &\quad + E\mathbf{h}_{\beta}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} E\mathbf{h}_{\theta} \\ &= D_1 C^{-1} E\mathbf{h}_{\theta} \\ I_{21} &= I_{12}^T \\ I_{22} &= E\mathbf{h}_{\theta}^T \{E\mathbf{h}\mathbf{h}^T - ER\mathbf{h}U^T(ERUU^T)^{-1}ERU\mathbf{h}^T\}^{-1} E\mathbf{h}_{\theta} \\ &= E\mathbf{h}_{\theta}^T C^{-1} E\mathbf{h}_{\theta} \end{aligned}$$

Note that

$$\text{ACov}(\hat{\beta}) = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$$

and

$$\begin{aligned} I_{11} - I_{12}I_{22}^{-1}I_{21} &= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} + D_1 C^{-1} D_1^T - D_1 C^{-1} E\mathbf{h}_{\theta} I_{22}^{-1} E\mathbf{h}_{\theta}^T C^{-1} D_1^T \\ &= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} + D_1 C^{-1} \left\{ C - E\mathbf{h}_{\theta} (E\mathbf{h}_{\theta}^T C^{-1} E\mathbf{h}_{\theta})^{-1} E\mathbf{h}_{\theta}^T \right\} C^{-1} D_1^T \\ &= ERU_{\beta}^T(ERUU^T)^{-1}ERU_{\beta} + \mathbf{Z}\mathbf{B}\mathbf{Z}^T. \end{aligned}$$

The desired result then follows. □

A.2 Lemma 2 and Proof

Lemma 2. For a symmetric, positive definite matrix $A_{m \times m}$ and a full rank matrix $G_{m \times p}$ with $p \leq m$,

$$A - G(G^T A^{-1} G)^{-1} G^T$$

is positive semi-definite.

Proof. $\text{rank}(G) = p$, so it has singular value decomposition as

$$G = O_{m \times m} \begin{bmatrix} D_{p \times p} \\ \mathbf{0} \end{bmatrix} N_{p \times p}^T$$

where O, N are orthorgnal and D is diagonal. Then

$$\begin{aligned} G^T A^{-1} G &= N [D \ \mathbf{0}] O^T A^{-1} O \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} N^T =: N [D \ \mathbf{0}] Q \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} N^T \\ &= N D Q_1 D N^T \end{aligned}$$

where $Q_{m \times m} = O^T A^{-1} O$, Q_1 is the first $q \times q$ diagonal block of Q , both invertible. In other words

$$O^T A^{-1} O = Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_2^T & Q_3 \end{bmatrix}$$

O, N are unitary and D is diagonal. Then

$$A - G(G^T A^{-1} G)^{-1} G^T = A - O \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} N^T N^{-T} (N D Q_1 D N^T)^{-1} N^{-1} N [D \ \mathbf{0}] O^T$$

$$= OQ^{-1}O^T - O \begin{bmatrix} Q_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} O^T$$

Since Q is symmetric and positive definite, it has a (unique) Cholesky decomposition $Q = LL^T$ where L is lower-triangular with positive diagonal entries. So we can write

$$Q = LL^T = \begin{bmatrix} L_{11} & \mathbf{0} \\ L_{12} & L_{22} \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{12}^T \\ \mathbf{0} & L_{22}^T \end{bmatrix} = \begin{bmatrix} L_{11}L_{11}^T = Q_1 & L_{11}L_{12}^T \\ L_{12}L_{11}^T & * \end{bmatrix}$$

where L_{11}, L_{22} are lower-triangular with positive diagonal entries (and hence invertible), and

$$Q_1^{-1} = L_{11}^{-T} L_{11}^{-1},$$

$$L^{-1} = \begin{bmatrix} L_{11}^{-1} & \mathbf{0} \\ * & L_{22}^{-1} \end{bmatrix}$$

For any m -dimensional vector $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$, where \mathbf{Z}_1 has dimension q and \mathbf{Z}_2 has dimension $m - q$,

$$\begin{aligned} & \mathbf{Z} \left(Q^{-1} - \begin{bmatrix} Q_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \mathbf{Z}^T \\ &= \mathbf{Z} \begin{bmatrix} L_{11}^{-T} & * \\ \mathbf{0} & L_{22}^{-T} \end{bmatrix} \begin{bmatrix} L_{11}^{-1} & \mathbf{0} \\ * & L_{22}^{-1} \end{bmatrix} \mathbf{Z}^T - \mathbf{Z} \begin{bmatrix} L_{11}^{-T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} L_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Z}^T \\ &= \|(L_{11}^{-1} \mathbf{Z}_1, *)\|_2^2 - \|(L_{11}^{-1} \mathbf{Z}_1, \mathbf{0})\|_2^2 \\ &\geq 0 \end{aligned}$$

So

$$Q^{-1} - \begin{bmatrix} Q_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

is positive semi-definite and so is $A - G(G^T A^{-1} G)^{-1} G^T$.

□

Appendix B

Additional Derivations and Numerical Results of Chapter 3

B.1 Derivations for relationships between the matrices \mathcal{C} and \mathcal{J}

For convenience and without loss of generality, we consider one dimensional x and z . Derivation for higher dimensions are the same. We will show that

$$\mathcal{C} = EU(\phi_0)U(\phi_0)^T = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{1\pi} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{2\pi} \\ \mathcal{C}_{1\pi}^T & \mathcal{C}_{2\pi}^T & \mathcal{C}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{22} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} & \mathcal{C}_{\pi\pi} \end{bmatrix} \quad (\text{B.1})$$

and

$$\begin{aligned}
\mathcal{J} &= -E \left\{ \frac{\partial \mathbf{U}(\phi_0)}{\partial \phi^T} \right\} = \begin{bmatrix} -E \left(R \frac{\partial^2 \log f_c}{\partial \beta \partial \beta^T} \right) & -E \left(R \frac{\partial^2 \log f_c}{\partial \beta \partial \alpha^T} \right) \\ -E \left(R \frac{\partial^2 \log f_c}{\partial \alpha \partial \beta^T} \right) & -E \left(R \frac{\partial^2 \log f_c}{\partial \alpha \partial \alpha^T} \right) \\ \mathbf{0} & -E \left\{ \frac{\partial}{\partial \alpha^T} \left(\frac{R}{\pi} - \frac{1-R}{1-\pi} \right) \frac{\partial \pi}{\partial \alpha} \right\} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{12}^T & \mathcal{I}_{22} \\ \mathbf{0} & \mathcal{I}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \\ \mathbf{0} & \mathcal{C}_{\pi\pi} \end{bmatrix}. \tag{B.2}
\end{aligned}$$

For a function h of y, x, z and r where r is binary, we use $\int h d\mu(y, r)$ to denote the integral with respect to y and summed over $r = 1$ and $r = 0$. Then we can compute

$$\begin{aligned}
\mathcal{I}_{11} &= -E \int \frac{\partial \mathbf{S}_1}{\partial \beta^T} f(y|x, z; \beta) d\mu(y, r) \\
&= -E \int r \frac{\partial^2 \log f_c(y|x, z; \beta, \alpha)}{\partial \beta \partial \beta^T} f(y|x, z; \beta) \pi(y, x; \alpha) d\mu(y, r) \\
&= -E \int r \frac{\partial^2 \log f_c(y|x, z; \beta, \alpha)}{\partial \beta \partial \beta^T} f_c(y|x, z; \beta, \alpha) \pi^*(x, z; \beta, \alpha) d\mu(y, r) \\
&= -E \int r \frac{\partial^2 \log f_c(y|x, z; \beta, \alpha)}{\partial \beta \partial \beta^T} f_c(y|x, z; \beta, \alpha) d\mu(y, r) \pi^*(x, z; \beta, \alpha) \\
&:= E\{\mathcal{I}_{c\beta\beta}(x, z) \pi^*(x, z; \beta, \alpha)\} \tag{B.3}
\end{aligned}$$

where the expectation is taken with respect to the marginal distribution $g(x, z)$; $\pi^*(x, z; \beta, \alpha) = f(y|x, z; \beta) \pi(y, x; \alpha) / f_c(y|x, z; \beta, \alpha) = \int f(y|x, z; \beta) \pi(y, x; \alpha) dy$, and

$$\begin{aligned}
\mathcal{I}_{c\beta\beta}(x, z) &= - \int r \frac{\partial^2 \log f_c(y|x, z; \beta, \alpha)}{\partial \beta \partial \beta^T} f_c(y|x, z; \beta, \alpha) d\mu(y, r) \\
&= - \int r \left\{ -\frac{1}{f_c^2} \left(\frac{\partial f_c}{\partial \beta} \right)^{\otimes 2} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \beta \partial \beta^T} \right\} f_c(y|x, z; \beta, \alpha) d\mu(y, r)
\end{aligned}$$

$$\begin{aligned}
&= \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\
&+ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} E \int r f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\
&= \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
\mathcal{C}_{11} &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\
&= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
&= E \{ \mathcal{I}_{c\beta\beta}(x, z) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \}.
\end{aligned} \tag{B.4}$$

So by (B.3) and (B.4), we have $\mathcal{I}_{11} = \mathcal{C}_{11}$.

Similarly, we compute

$$\begin{aligned}
\mathcal{I}_{1\pi} = \mathcal{I}_{12} &= -E \int \frac{\partial \mathbf{S}_1}{\partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\
&= -E \int r \frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\
&:= E \{ \mathcal{I}_{c\beta\alpha}(x, z) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \}
\end{aligned} \tag{B.5}$$

where

$$\begin{aligned}
\mathcal{I}_{c\beta\alpha}(x, z) &= - \int r \frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\
&= - \int r \left\{ -\frac{1}{f_c^2} \frac{\partial f_c}{\partial \boldsymbol{\beta}} \frac{\partial f_c}{\partial \boldsymbol{\alpha}^T} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} \right\} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\
&= \int r \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\
&+ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} \int r f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r)
\end{aligned}$$

$$= \int r \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r).$$

$$\begin{aligned} \mathcal{C}_{12} &= E \int r \frac{\partial \log f_c}{\partial \boldsymbol{\beta}} \frac{\partial \log f_c}{\partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= E\{\mathcal{I}_{c\beta\alpha}(x, z) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\}. \end{aligned} \quad (\text{B.6})$$

At the same time, if we differentiate both sides of $E(\mathbf{S}_1|x, z) = \mathbf{0}$ with respect to $\boldsymbol{\alpha}$, we have

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\alpha}^T} \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &= \int r \frac{\partial^2 \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &\quad + \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} f(y|x, z; \boldsymbol{\beta}) \frac{\partial \pi(y, x; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} d\mu(y, r). \end{aligned}$$

Taking expectation with respect to the marginal distribution $g(x, z)$, we get

$$\mathbf{0} = -\mathcal{I}_{1\pi} + E \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \frac{\partial \pi(y, x; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) d\mu(y, r) \quad (\text{B.7})$$

Note that the first term above is equal to $-\mathcal{I}_{12}$. Since $r^2 = r$ and $r(1-r) = 0$, we can rewrite the second term as

$$\begin{aligned} &E \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \left\{ \frac{r}{\pi(y, x; \boldsymbol{\alpha})} - \frac{1-r}{1-\pi(y, x; \boldsymbol{\alpha})} \right\} \frac{\partial \pi(y, x; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} \\ &\quad \cdot f(y|x, z; \boldsymbol{\beta}) \pi(y, x, \boldsymbol{\alpha}) d\mu(y, r) \\ &= E \int \mathbf{S}_1 \mathbf{S}_\pi^T f(y|x, z; \boldsymbol{\beta}) \pi(y, x, \boldsymbol{\alpha}) d\mu(y, r) \\ &= E(\mathbf{S}_1 \mathbf{S}_\pi^T) = \mathcal{C}_{1\pi}. \end{aligned} \quad (\text{B.8})$$

Thus by (B.5, B.6, B.7, B.8), we get

$$\mathcal{I}_{1\pi} = \mathcal{I}_{12} = \mathcal{C}_{1\pi} = \mathcal{C}_{12}. \quad (\text{B.9})$$

Similarly,

$$\begin{aligned} \mathcal{I}_{22} &= -E \int \frac{\partial \mathcal{S}_2}{\partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &= -E \int r \frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &:= E\{\mathcal{I}_{c\alpha\alpha}(x, z) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\} \end{aligned} \quad (\text{B.10})$$

and

$$\begin{aligned} \mathcal{I}_{c\alpha\alpha}(x, z) &= - \int r \frac{\partial^2 \log f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\ &= - \int r \left\{ -\frac{1}{f_c^2} \left(\frac{\partial f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} + \frac{1}{f_c} \frac{\partial^2 f_c}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \right\} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\ &= \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) + \frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \int r f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \\ &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r). \end{aligned}$$

We also have

$$\begin{aligned} \mathcal{C}_{22} &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &= E \int r \left(\frac{\partial \log f_c}{\partial \boldsymbol{\alpha}} \right)^{\otimes 2} f_c(y|x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) d\mu(y, r) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha}) \\ &= E\{\mathcal{I}_{c\alpha\alpha}(x, z) \pi^*(x, z; \boldsymbol{\beta}, \boldsymbol{\alpha})\}. \end{aligned} \quad (\text{B.11})$$

Differentiate both sides of $E(\mathbf{S}_2|x, z) = \mathbf{0}$ with respect to $\boldsymbol{\alpha}$, and we have

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\alpha}^T} \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &= \int r \frac{\partial^2 \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) \pi(y, x; \boldsymbol{\alpha}) d\mu(y, r) \\ &\quad + \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} f(y|x, z; \boldsymbol{\beta}) \frac{\partial \pi(y, x; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} d\mu(y, r) \end{aligned}$$

and taking expectation with respect to $g(x, z)$,

$$\begin{aligned} \mathbf{0} &= -\mathcal{I}_{22} + E \int r \frac{\partial \log f_c(y|x, z, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \frac{\partial \pi(y, x; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^T} f(y|x, z; \boldsymbol{\beta}) d\mu(y, r) \\ &= -\mathcal{I}_{22} + E(\mathbf{S}_2 \mathbf{S}_\pi^T) = -\mathcal{I}_{21} + \mathcal{C}_{2\pi}. \end{aligned} \tag{B.12}$$

Thus by (B.10, B.11, B.12), we see that

$$\mathcal{I}_{22} = \mathcal{C}_{2\pi} = \mathcal{C}_{22}. \tag{B.13}$$

Lastly, by maximum likelihood for logistic regression we have $-E \left\{ \frac{\partial \mathbf{S}_\pi(\phi_0)}{\partial \boldsymbol{\alpha}^T} \right\} = E(\mathbf{S}_\pi \mathbf{S}_\pi^T)$,

so

$$\mathcal{I}_{\pi\pi} = \mathcal{C}_{\pi\pi}. \tag{B.14}$$

B.2 Calculation of the asymptotic variance of the EL estimator

We denote

$$\mathcal{C}^{-1} = D = \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix}, \quad (\text{B.15})$$

and then by $\mathcal{C}D = I$, we have $\mathcal{C}_{11}D^{11} + \mathcal{C}_{12}D^{21} + \mathcal{C}_{12}D^{31} = I$, $\mathcal{C}_{11}D^{12} + \mathcal{C}_{12}D^{22} + \mathcal{C}_{12}D^{32} = \mathbf{0}$, $\mathcal{C}_{11}D^{13} + \mathcal{C}_{12}D^{23} + \mathcal{C}_{12}D^{33} = \mathbf{0}$, and therefore

$$\begin{aligned} V^{11} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{12}^T \\ \mathbf{0} \end{bmatrix} \\ &= \begin{bmatrix} I - \mathcal{C}_{12}D^{31} & -\mathcal{C}_{12}D^{32}, & -\mathcal{C}_{12}D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} \\ \mathcal{C}_{12}^T \\ \mathbf{0} \end{bmatrix} \\ &= \mathcal{C}_{11} - \mathcal{C}_{12}(D^{31}\mathcal{C}_{11} + D^{32}\mathcal{C}_{12}^T) \\ &= \mathcal{C}_{11} + \mathcal{C}_{12}D^{33}\mathcal{C}_{12}^T. \end{aligned}$$

We can further compute D^{33} as:

$$D^{33} = \left(\mathcal{C}_{\pi\pi} - \begin{bmatrix} \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \end{bmatrix} \right)^{-1}.$$

Note that $\begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} = I$, so

$$\begin{bmatrix} \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^T & \mathcal{C}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \end{bmatrix} = \mathcal{C}_{22},$$

and therefore

$$D^{33} = (\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}. \quad (\text{B.16})$$

Thus, finally we obtain

$$V^{11} = \mathcal{C}_{11} + \mathcal{C}_{12}(\mathcal{C}_{\pi\pi} - \mathcal{C}_{22})^{-1}\mathcal{C}_{12}^T.$$

Similarly, V^{12} is the $(1, 2)$ block of $\mathcal{I}^T \mathcal{C}^{-1} \mathcal{I}$, so

$$\begin{aligned} V^{12} &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} D^{11} & D^{12} & D^{13} \\ D^{21} & D^{22} & D^{23} \\ D^{31} & D^{32} & D^{33} \end{bmatrix} \begin{bmatrix} \mathcal{C}_{12} \\ \mathcal{C}_{22} \\ \mathcal{C}_{\pi\pi} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ I \end{bmatrix} = \mathbf{0}. \end{aligned}$$

B.3 The rank of CL estimating equations for Studies 1 and 3

With the models in Simulation Study 1, both the regression model and π model are in logistic form, so as discussed in Scott and Wild (2011), the conditional probability $p(Y =$

$1|X, Z, R = 1)$ is also a logistic form, with an offset term $\omega_i = \log(\pi(y = 1, z)/\pi(y = 0, z))$.

Thus the conditional log-likelihood is

$$l_c(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^N r_i [y_i \log\{\text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\} \\ + (1 - y_i) \log\{1 - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\}]$$

and

$$\frac{\partial l_c}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N r_i \{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\} (1, x_i, z_i)^T,$$

$$\frac{\partial l_c}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^N r_i \{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\} \frac{\partial \omega_i}{\partial \boldsymbol{\alpha}}$$

When we use the “sat2” selection model, we have

$$\begin{aligned} \frac{\partial \omega_i}{\partial \boldsymbol{\alpha}} &= \frac{\partial}{\partial \boldsymbol{\alpha}} [\log\{\text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i)\}] - \frac{\partial}{\partial \boldsymbol{\alpha}} [\log\{\text{expit}(\alpha_c + \alpha_x x_i)\}] \\ &= \{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i)\} (1, 1, x_i, x_i)^T \\ &\quad - \{1 - \text{expit}(\alpha_c + \alpha_x x_i)\} (1, 0, x_i, 0)^T \\ &= \begin{pmatrix} \{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i)\} - \{1 - \text{expit}(\alpha_c + \alpha_x x_i)\} \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i) \\ x_i [\{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i)\} - \{1 - \text{expit}(\alpha_c + \alpha_x x_i)\}] \\ x_i \{1 - \text{expit}(\alpha_c + \alpha_y + \alpha_x x_i + \alpha_{yx} x_i)\} \end{pmatrix}. \end{aligned} \quad (\text{B.17})$$

As Z is a continuous variable, it is easy to see that $\partial \omega_i / \partial \boldsymbol{\alpha}$ in (B.17) is a full rank vector (no row of it is a linear combination of other rows).

However, when we use the “sat1” selection model where $\pi(y, x; \alpha) = \pi(y, v(z); \alpha)$, with

$v(x)$ a binary function of x , then

$$\begin{aligned}
\frac{\partial \omega_i}{\partial \mathbf{a}} &= v_i \begin{pmatrix} -\text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) + \text{expit}(\alpha_c + \alpha_v) \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) \\ -\text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) + \text{expit}(\alpha_c + \alpha_v) \\ 1 - \text{expit}(\alpha_c + \alpha_y + \alpha_v + \alpha_{yv}) \end{pmatrix} \\
&+ \begin{pmatrix} -\text{expit}(\alpha_c + \alpha_y) + \text{expit}(\alpha_c) \\ 1 - \text{expit}(\alpha_c + \alpha_y) \\ 0 \\ 0 \end{pmatrix} \\
&=: v_i(a_1, a_2, a_1, a_2)^T + (1 - v_i)(b_1, b_2, 0, 0)^T \\
&=: v_i \mathbf{a} + (1 - v_i) \mathbf{b}
\end{aligned}$$

where \mathbf{a}, \mathbf{b} are constant vectors. Thus the information matrix can be written as

$$E \left(\frac{\partial \log f_c}{\partial \boldsymbol{\phi}} \right) \left(\frac{\partial \log f_c}{\partial \boldsymbol{\phi}^T} \right) = E [r_i \{y_i - \text{expit}(\omega_i + \beta_c + \beta_x x_i + \beta_z z_i)\}^2 \mathbf{u}_i \mathbf{u}_i^T]$$

where

$$\begin{aligned}
\mathbf{u}_i &= (1, x_i, z_i, a_1 v_i + b_1(1 - v_i), a_2 v_i + b_2(1 - v_i), a_1 v_i, a_2 v_i)^T \\
&= \begin{bmatrix} 1 & 0 & 0 & b_1 & b_2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 - b_1 & a_2 - b_2 & a_1 & a_2 \end{bmatrix}^T \begin{bmatrix} 1 \\ x_i \\ z_i \\ v_i \end{bmatrix} := U \times (1, x_i, z_i, v_i)^T
\end{aligned}$$

where U is a 7×4 constant matrix. Thus $E (\partial \log f_c / \partial \boldsymbol{\phi}) (\partial \log f_c / \partial \boldsymbol{\phi})^T$ has dimension 7×7 but rank 4.

B.4 Additional simulation studies

B.4.1 Simulation study 3

This study involves a binary covariate X and continuous covariate Z , which are correlated. We consider a phase 1 sample of 10,000 subjects with the complete data generated as Section 3.3.1. In phase 2, we randomly sample $n_1 = 150$ subjects from the cases with $Y_i = 1$, and $n_0 = 150$ from the subjects with $Y_i = 0$; the X_i are discarded for all other subjects and marked as unobserved.

This is a case of basic stratified sampling (BSS) with the phase 2 sampling depending only on the observed values of Y . The marginal sampling probability for $Y = 1$ cases is $p_1 = 150/N_1$ and for $Y = 0$ cases is $p_0 = 150/N_0$ but the R_i are not independent as for variable probability sampling (VPS). We can nevertheless use the VPS estimating equations and likelihoods, which are asymptotically valid under BSS; we do this, although finite sample adjustments for BSS could be made (e.g. Lawless et al. 1999). Under VPS we would use a logistic regression model for the sampling probabilities:

$$P(R = 1|y) = \pi_{est}(y; \boldsymbol{\alpha}) = \text{expit}(\alpha_c + \alpha_y y), \quad (\text{B.18})$$

but in the present case the design probabilities p_0, p_1 are random and not fixed, since they depend on N_0 and N_1 . We denote estimates obtained using these design probabilities with the suffix *est* in Table B.1. It is possible, however, to increase efficiency of estimation by using a stratified pseudo VPS sampling model that conditions on observed z values, similar to calibration or post-stratification in sampling contexts. We consider two such models, referred to with the suffixes *sat1* and *sat2* in Table B.1. For *sat1* we use a binary covariate $v = I(Z > 0.5)$ and the same π model as (3.8) and the *sat2* model uses the model as (3.9). Note that working models (8) and (9) both include the true phase 2 sampling model (7) as

special cases.

We also considered two pseudo empirical likelihood (PEL) estimators, where the α parameters in models (B.18), (3.8) and (3.9) are first estimated by maximum likelihood from $\mathbf{S}_\pi(\alpha) = 0$ and then fixed in the estimating function $\mathbf{U}(\phi) = \mathbf{U}(\beta, \hat{\alpha}_{\text{ML}})$. This EL procedure is slightly easier to implement since the estimating function $\mathbf{S}_\pi(\hat{\alpha}_{\text{ML}})$ equals zero. Such estimators have been considered by others such as Qin et al. (2009) and Xie and Zhang (2017).

We mention that in this example the estimating equations \mathbf{S}_1 and \mathbf{S}_2 are not linearly independent. Take the π_{sat1} model, for example; then $\dim(\beta) = 3$ and $\dim(\alpha) = 4$ so the dimension of $(\mathbf{S}_1^T, \mathbf{S}_2^T)^T$ is 7. However in Appendix Section A.3 we show that the actual rank of these 7 estimating equations is 4. Therefore we use here only the first element of \mathbf{S}_2 for the EL estimator. This phenomenon is an example of the well known fact that β and α are not identifiable from the conditional likelihood $l_c(\beta, \alpha)$ alone in this setting.

In Table B.1, we compare the performance of CML, SW and EL estimators based on 500 simulations, using each of the three π models (7) - (9). The EL0 and PEL estimator with each π model are asymptotically equivalent to the corresponding EL estimator so are omitted; their finite sample performances are close to those of the EL estimators. We show empirical standard deviations and average standard errors for each estimator; standard errors are obtained by estimating asymptotic covariance matrices with sample covariance matrices evaluated at estimates of ϕ . These are labelled empirical and estimated standard error (SE) in the table and they are seen to be close in value. In this case, CML performs about as well as the EL and SW methods. A substantial efficiency gain for estimation of β_Z , the coefficient for the covariate that is known for all individuals, occurs when the stratified selection model (8) is used instead of (7) for the EL and SW estimators. A big increase in efficiency for CML and small further increases in efficiency for EL and SW result from using the more highly

Table B.1: Simulation results for Study 3.

Method	Mean (Empirical SE)[Estimated SE]		
	$\beta_c (\beta_{c0} = -2.8)$	$\beta_z (\beta_{z0} = 0.5)$	$\beta_x (\beta_{x0} = 1)$
CML-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
CML-sat1	-2.815 (0.115)[0.122]	0.524 (0.198)[0.200]	1.020 (0.239)[0.250]
CML-sat2	-2.814 (0.113)[0.120]	0.524 (0.124)[0.124]	1.021 (0.239)[0.250]
EL-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
EL-sat1	-2.814 (0.116)[0.122]	0.514 (0.130)[0.134]	1.020 (0.239)[0.249]
EL-sat2	-2.814 (0.114)[0.120]	0.520 (0.122)[0.123]	1.019 (0.240)[0.250]
SW-est	-2.813 (0.117)[0.123]	0.522 (0.247)[0.257]	1.018 (0.239)[0.250]
SW-sat1	-2.814 (0.116)[0.122]	0.515 (0.131)[0.130]	1.020 (0.239)[0.249]
SW-sat2	-2.814 (0.113)[0.120]	0.518 (0.121)[0.123]	1.018 (0.239)[0.250]

stratified model (9).

B.4.2 Simulation study 4

In Study 4, we again simulate a normal linear regression model, but now with X and Z both continuous. We let X, Z follow a bivariate normal distribution with zero means, variances one and correlation $\rho = 0.5$. The response model is $Y \sim \mathcal{N}(0.5X + Z, 1)$, and so $\beta_0 = (0, 0.5, 1)$. The phase 1 sample size is $N = 500$ and the phase 2 sampling probability model is $P(R = 1|y, z) = \text{expit}(-1 + 0.5y + 0.5z)$, resulting in about 30% of subjects being selected in phase 2. In this case, we have the conditional likelihood

$$f_c(y|x, z; \beta, \alpha) = \frac{\exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2 / (2\sigma^2)\} \text{expit}(\alpha_c + \alpha_y y + \alpha_z z)}{\int \exp\{-(y - \beta_c - \beta_x x - \beta_z z)^2 / (2\sigma^2)\} \text{expit}(\alpha_c + \alpha_y y + \alpha_z z) dy}. \quad (\text{B.19})$$

Table B.2: Simulation results for Study 4.

Method	Mean (Empirical SE)[Estimated SE]			
	$\beta_c (\beta_{c0} = 0)$	$\beta_z (\beta_{z0} = 0.5)$	$\beta_x (\beta_{x0} = 1)$	$\sigma (\sigma_0 = 1)$
CML0	0.006 (0.102)[0.106]	0.494 (0.091)[0.092]	1.000 (0.093)[0.091]	0.985 (0.060)[0.062]
CML-est	0.008 (0.081)[0.093]	0.493 (0.075)[0.091]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
CML-sat	0.005 (0.080)[0.092]	0.498 (0.076)[0.085]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
EL-est	0.011 (0.084)[0.087]	0.489 (0.089)[0.090]	0.995 (0.093)[0.088]	0.980 (0.062)[0.060]
EL-sat	0.008 (0.082)[0.085]	0.499 (0.075)[0.081]	0.993 (0.092)[0.088]	0.979 (0.062)[0.060]
SW-est	0.005 (0.074)[0.086]	0.498 (0.076)[0.082]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]
SW-sat	0.005 (0.074)[0.086]	0.498 (0.076)[0.082]	1.000 (0.091)[0.089]	0.985 (0.061)[0.061]

We consider the two phase 2 selection models

$$\pi_{est}(y, z; \boldsymbol{\alpha}) = P(R = 1|y, z) = \text{expit}(\alpha_c + \alpha_y y + \alpha_z z) \quad (\text{B.20})$$

$$\pi_{sat}(y, z; \boldsymbol{\alpha}) = P(R = 1|y, z) = \text{expit}(\alpha_c + \alpha_y y + \alpha_z z + \alpha_{yz} yz) \quad (\text{B.21})$$

for CML, SW and EL estimation. The performances of the estimators in 100 simulations are compared in Table B.2. Once again we find that with the most highly stratified model (B.21), the three estimators have almost identical empirical standard errors for β_z , and that EL and SW estimators are slightly more efficient for estimation of β_c .

Appendix C

Additional Derivations and Numerical Results of Chapter 4

C.1 Proofs of (4.6)

The Lagrangian of the corresponding optimization is

$$\mathcal{L} = \sum_{i=1}^m (\log f_{c,i} + \log p_i) + m\boldsymbol{\lambda}^T \sum_{i=1}^m p_i \mathbf{u}_i(\boldsymbol{\beta}) + \mu \left(\sum_{i=1}^m p_i - 1 \right),$$

At the solution $\hat{\boldsymbol{\beta}}_{\text{EL0-1}}$ and \hat{p}_i we have $\partial \mathcal{L} / \partial p_i = 0$ and $\partial \mathcal{L} / \partial \boldsymbol{\beta} = 0$, which yields

$$\hat{p}_i = 1 / [m \{1 - \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{\text{EL0-1}})\}]$$

and

$$0 = \sum_{i=1}^m s_{c,i} + m\boldsymbol{\lambda}^T \sum_{i=1}^m \hat{p}_i \partial \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{\text{EL0-1}}) / \partial \boldsymbol{\beta},$$

$$0 = \sum_{i=1}^m \hat{p}_i \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{\text{EL0-1}}).$$

Applying the mean value theorem to the last two equations around (β_0) and 0, we have

$$\begin{aligned}
0 &= \begin{pmatrix} 1/m \sum_{i=1}^m \mathbf{s}_{c,\beta,i}(\beta_0) \\ 1/m \sum_{i=1}^m \mathbf{u}_i(\beta_0) \end{pmatrix} \\
&+ \begin{pmatrix} 1/m \sum_{i=1}^m \frac{\partial \mathbf{s}_{c,\beta,i}(\beta_0)}{\partial \beta} & 1/m \sum_{i=1}^m \frac{\partial \mathbf{u}_i(\hat{\beta}_{\text{EL0-1}})/\partial \beta}{1 - \hat{\lambda}^T \mathbf{u}_i(\hat{\beta}_{\text{EL0-1}})} \\ 1/m \sum_{i=1}^m \frac{\partial \mathbf{u}_i(\bar{\beta})/\partial \beta}{1 - \bar{\lambda}^T \mathbf{u}_i(\hat{\beta}_{\text{EL0-1}})} & 1/m \sum_{i=1}^m \frac{\mathbf{u}_i(\bar{\beta}) \mathbf{u}_i(\hat{\beta}_{\text{EL0-1}})}{1 - \bar{\lambda}^T \mathbf{u}_i(\hat{\beta}_{\text{EL0-1}})} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{\text{EL0-1}} - \beta_0 \\ \hat{\lambda} \end{pmatrix} \\
&:= \mathbf{v}_m + J_m
\end{aligned}$$

for some $\bar{\beta}$ between β_0 and $\hat{\beta}_{\text{SEL1}}$ and some $\bar{\lambda}$ between $\mathbf{0}$ and $\hat{\lambda}$. Then we have

$$\sqrt{m} \mathbf{v}_m \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega} \end{pmatrix},$$

with $\mathbf{S} = E\{R \mathbf{s}_{c,\beta}(\beta_0) \mathbf{s}_{c,\beta}(\beta_0)^T\}$, $\mathbf{\Omega} = E\{R \mathbf{u}(\beta_0) \mathbf{u}(\beta_0)^T\}$. The off-diagonal block is zero as

$$\begin{aligned}
E\{R \mathbf{s}_{c,\beta}(\beta_0) \mathbf{u}(\beta_0)^T | R = 1\} &= E[E\{R \mathbf{s}_{c,\beta}(Y, X, Z; \beta_0) \mathbf{u}(X, Z; \beta_0, \alpha_0, \theta_0) | X, Z\}] \\
&= E[E\{R \mathbf{s}_{c,\beta}(Y | X, Z; \beta_0)\} \mathbf{u}(X, Z; \beta_0, \alpha_0, \theta_0)]
\end{aligned}$$

and we know $E\{R \mathbf{s}_{c,\beta}(Y | X, Z; \beta_0)\} = 0$.

Denoting $\mathbf{J} = E\{R \partial \mathbf{u}(\beta_0) / \partial \beta\}$, we have

$$J_m \xrightarrow{p} J := \begin{pmatrix} E\{R \frac{\partial \mathbf{s}(\beta_0)}{\partial \beta^T}\} & E\{R \frac{\partial \mathbf{u}(\beta_0)}{\partial \beta}\} \\ E\{R \frac{\partial \mathbf{u}(\beta_0)}{\partial \beta^T}\} & E\{R \mathbf{u}(\beta_0) \mathbf{u}(\beta_0)^T\} \end{pmatrix} = \begin{pmatrix} -\mathbf{S} & \mathbf{J} \\ \mathbf{J}^T & \mathbf{\Omega} \end{pmatrix},$$

thus

$$\sqrt{m} \begin{pmatrix} \hat{\beta}_{\text{EL0-1}} - \beta_0 \\ \hat{\lambda} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \mathbf{J}^{-1} \Sigma \mathbf{J}^{-T}).$$

and $\text{ACov}\{\sqrt{m}(\hat{\beta}_{\text{EL0-1}} - \beta_0)\}$ equals the upper left block of $\mathbf{J}^{-1} \Sigma \mathbf{J}^{-T}$, which is $\Sigma_1 = (\mathbf{S} + \mathbf{J} \Omega \mathbf{J}^T)^{-1}$.

C.2 Equivalence of the EL0-1 and EL1 estimator

For the EL1 estimator, its asymptotic covariance can be easily derived from the result of empirical likelihood for general estimating equations (Qin and Lawless 1994) as

$$\left[E \left\{ \frac{\partial \mathbf{g}(\beta_0)}{\partial \beta} \right\} E \{ \mathbf{g}(\beta_0) \mathbf{g}(\beta_0)^T \}^{-1} E \left\{ \frac{\partial \mathbf{g}(\beta_0)}{\partial \beta^T} \right\} \right]^{-1}$$

where \mathbf{g} is the set of all estimating functions, namely $\mathbf{g} = R(\mathbf{s}_{c,\beta}, \mathbf{u})$. Then

$$\text{ACov}\{\sqrt{m}(\hat{\beta}_{\text{EL1}} - \beta_0)\} = (\mathbf{S} + \mathbf{J} \Omega \mathbf{J}^T)^{-1} = \text{ACov}\{\sqrt{m}(\hat{\beta}_{\text{EL0-1}} - \beta_0)\}$$

. Therefore, EL0-1 and EL1 estimator is asymptotically equivalent.

C.3 Equivalence of the EL0-2 and EL2 estimator

Similarly as in Section C.1, we note that $\hat{\beta}_{\text{EL0-2}}$ and the Lagrange multiplier $\hat{\lambda}$ satisfy

$$\sum_{i=1}^n \left\{ r_i \mathbf{s}_{c,\beta}(\hat{\beta}_{\text{EL0-2}}, \alpha_0) + \frac{r_i \partial \mathbf{u}_i(\hat{\beta}_{\text{EL0-2}}, \alpha_0, \hat{\theta}) / \partial \beta^T}{1 + \hat{\lambda}^T \mathbf{u}_i(\hat{\beta}_{\text{EL0-2}}, \alpha_0, \hat{\theta})} \hat{\lambda} \right\} = \mathbf{0}$$

$$\sum_{i=1}^n \left\{ \frac{r_i \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{\text{ELO-2}}, \boldsymbol{\alpha}_0, \hat{\boldsymbol{\theta}})}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i(\hat{\boldsymbol{\beta}}_{\text{ELO-2}}, \boldsymbol{\alpha}_0, \hat{\boldsymbol{\theta}})} \right\} = \mathbf{0},$$

and $\sum_{i=1}^n \{ \mathbf{h}(Y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}) \} = \mathbf{0}.$

By a first-order Taylor expansion,

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} r_i \mathbf{s}_{c,\beta,i}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) \\ r_i \mathbf{u}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) \\ \mathbf{h}(\boldsymbol{\theta}_0) \end{pmatrix} + \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \frac{r_i \partial \mathbf{s}_{c,\beta,i}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)}{\partial \boldsymbol{\beta}^T} & \frac{\partial \mathbf{u}_i / \partial \boldsymbol{\beta}^T}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i} & \mathbf{0} \\ \frac{r_i \partial \mathbf{u}_i / \partial \boldsymbol{\beta}^T}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i} & \frac{\mathbf{u}_i \mathbf{u}_i^T}{(1 + \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i)^2} & \frac{r_i \partial \mathbf{u}_i / \partial \boldsymbol{\theta}^T}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{u}_i} \\ \mathbf{0} & \mathbf{0} & \frac{\partial \mathbf{h}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{ELO-2}} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0 \end{pmatrix} + o_p(1) \\ &:= v_n + G_n \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{SEL}} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0 \end{pmatrix} + o_p(1), \end{aligned}$$

and by CLT and law of large numbers,

$$\sqrt{n} \mathbf{v}_n \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{S} & \mathbf{0} & \mathbf{U}^T \\ \mathbf{0} & \boldsymbol{\Omega} & \mathbf{V}^T \\ \mathbf{U} & \mathbf{V} & \mathbf{W} \end{pmatrix},$$

where $V = E(hu^T)$ and $W = E(hh^T)$; G_n/n converges in probability to constant matrix

$$G := \begin{pmatrix} -\mathbf{S} & \mathbf{J}^T & \mathbf{0} \\ \mathbf{J} & \mathbf{\Omega} & \mathbf{H} \\ \mathbf{0} & \mathbf{0} & -\tilde{\mathbf{W}} \end{pmatrix} = \begin{pmatrix} E\{R\partial \mathbf{s}_{c,\beta}/\partial \boldsymbol{\beta}^T\} & E\{R\partial \mathbf{u}/\partial \boldsymbol{\beta}^T\} & \mathbf{0} \\ E\{R\partial \mathbf{u}/\partial \boldsymbol{\beta}^T\} & E\{R\mathbf{u}\mathbf{u}^T\} & E\{R\partial \mathbf{u}/\partial \boldsymbol{\theta}^T\} \\ \mathbf{0} & \mathbf{0} & E\{\partial \mathbf{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T\} \end{pmatrix},$$

thus

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{EL0-2}} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0 \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, G^{-1} \boldsymbol{\Sigma} G^{-T}).$$

It is to be noted that

$$\begin{aligned} E\{R\partial \mathbf{u}(X, Z)/\partial \boldsymbol{\theta}^T\} &= E \left[R \frac{\partial}{\partial \boldsymbol{\theta}^T} E \left\{ \frac{\mathbf{h}(Y, X)}{\pi(Y, X)} | X, Z, R = 1 \right\} \right] \\ &= E \left[R E \left\{ \frac{\partial \mathbf{h}(Y, X)/\partial \boldsymbol{\theta}^T}{\pi(Y, X)} | X, Z, R = 1 \right\} \right] \\ &= E \left[\frac{\partial \mathbf{h}(Y, X)}{\partial \boldsymbol{\theta}^T} \right] \end{aligned}$$

therefore $\mathbf{H} = \tilde{\mathbf{W}}$, and both are symmetric.

$$\begin{aligned} G^{-1} &= \begin{pmatrix} -\mathbf{S} & \mathbf{J}^T & \mathbf{0} \\ \mathbf{J} & \mathbf{\Omega} & -\mathbf{H} \\ \mathbf{0} & \mathbf{0} & -\mathbf{H} \end{pmatrix}^{-1} = \begin{pmatrix} \begin{pmatrix} -\mathbf{S} & \mathbf{J}^T \\ \mathbf{J} & \mathbf{\Omega} \end{pmatrix}^{-1} & \begin{pmatrix} -\mathbf{S} & \mathbf{J}^T \\ \mathbf{J} & \mathbf{\Omega} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ -\mathbf{H} \end{pmatrix} \mathbf{H}^{-1} \\ \mathbf{0} & \mathbf{0} & -\mathbf{H}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} -(\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} & (\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} \mathbf{J}^T \mathbf{\Omega}^{-1} & -(\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} \mathbf{J} \mathbf{\Omega}^{-1} \\ * & * & * \\ \mathbf{0} & \mathbf{0} & * \end{pmatrix} \end{aligned}$$

$$= (\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} \begin{pmatrix} -\mathbf{I} & \mathbf{J}^T \mathbf{\Omega}^{-1} & -\mathbf{J}^T \mathbf{\Omega}^{-1} \\ * & * & * \\ \mathbf{0} & \mathbf{0} & * \end{pmatrix},$$

Hence

$$\begin{aligned} & \text{ACov}\{\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{EL0-2}} - \boldsymbol{\beta}_0)\} \\ &= (\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} \begin{pmatrix} -\mathbf{I} & \mathbf{J}^T \mathbf{\Omega}^{-1} & -\mathbf{J}^T \mathbf{\Omega}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{S} & \mathbf{0} & \mathbf{U}^T \\ \mathbf{0} & \mathbf{\Omega} & \mathbf{V}^T \\ \mathbf{U} & \mathbf{V} & \mathbf{W} \end{pmatrix} \begin{pmatrix} -\mathbf{I} \\ \mathbf{\Omega}^{-1} \mathbf{J} \\ \mathbf{\Omega}^{-1} \mathbf{J} \end{pmatrix} (\mathbf{S} + \mathbf{J}^T \mathbf{\Omega}^{-1} \mathbf{J})^{-1} \\ &= (\mathbf{S} + \mathbf{J} \mathbf{\Omega}^{-1} \mathbf{J}^T)^{-1} \{ \mathbf{S} - \mathbf{J} \mathbf{\Omega}^{-1} \mathbf{U} - \mathbf{U}^T \mathbf{\Omega}^{-1} \mathbf{J}^T + \mathbf{J} \mathbf{\Omega}^{-1} (\mathbf{\Omega} - \mathbf{V} - \mathbf{V}^T + \mathbf{W}) \mathbf{\Omega}^{-1} \mathbf{J}^T \} \\ & \quad \cdot (\mathbf{S} + \mathbf{J} \mathbf{\Omega}^{-1} \mathbf{J}^T)^{-1}. \end{aligned}$$

There is no direct comparison between $\text{ACov}\{\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{EL0-1}} - \boldsymbol{\beta}_0)\}$ and $\text{ACov}\{\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{EL0-2}} - \boldsymbol{\beta}_0)\}$.

However, we can show the equivalence of $\hat{\boldsymbol{\beta}}_{\text{EL0-2}}$ and $\hat{\boldsymbol{\beta}}_{\text{EL2}}$, and easily derive that both of them are more efficient than the classic CML estimator. The EL2 estimator is defined through

$$\begin{aligned} & \max_{\boldsymbol{\beta}, p_1, \dots, p_n} \prod_{i=1}^n p_i \quad \text{subject to } p_i \geq 0, \sum_{i=1}^n p_i = 1, \\ & \text{and } \sum_{i=1}^n p_i \begin{pmatrix} \mathbf{u}(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}_0, \hat{\boldsymbol{\theta}}) \\ \mathbf{s}_{c, \boldsymbol{\beta}}(Y_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_0) \end{pmatrix} = \mathbf{0}, \end{aligned} \tag{C.1}$$

and we omit the parameter $\boldsymbol{\alpha}_0$ as it is fixed. By Lagrange multipliers method as in Qin and Lawless (1994), one can write the profile likelihood as

$$l(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = - \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta}, \boldsymbol{\theta})\}$$

where $g_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = (R_i \mathbf{u}(X_i, Z_i, \boldsymbol{\beta}, \boldsymbol{\theta})^T, R_i \mathbf{s}_{c,\beta}(Y_i, X_i, Z_i, \boldsymbol{\beta})^T)^T$. Write

$$\begin{aligned} Q_{1n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}) &= -\frac{1}{n} \frac{\partial l}{\partial \boldsymbol{\lambda}} = \frac{1}{n} \sum_{i=1}^n \frac{g_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta}, \boldsymbol{\theta})}, \\ Q_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}) &= -\frac{1}{n} \frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{\lambda}^T \partial g_i(\boldsymbol{\beta}, \boldsymbol{\theta}) / \partial \boldsymbol{\beta}}{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta}, \boldsymbol{\theta})}, \\ Q_{3n}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta}). \end{aligned}$$

Let $\hat{\boldsymbol{\lambda}}$ denote the Lagrange multiplier that satisfies the constraint

$$\sum_{i=1}^n \frac{g_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta}, \boldsymbol{\theta})} = 0,$$

by Z-estimator theory, $\hat{\boldsymbol{\beta}}_{\text{EL2}}$ and $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ are both consistent estimators, and by general empirical likelihood theory, $\hat{\boldsymbol{\lambda}} = O_p(n^{-1/2})$. An application of a first-order Taylor expansion around $(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)$ yields

$$\begin{aligned} \mathbf{0} &= Q_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0) + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}}_{\text{EL2}} - \boldsymbol{\beta}_0) + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\lambda}^T} \hat{\boldsymbol{\lambda}} \\ &\quad + \frac{\partial Q_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) + O_p(n^{-1}) \\ \mathbf{0} &= Q_{2n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0) + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}^T} (\hat{\boldsymbol{\beta}}_{\text{EL2}} - \boldsymbol{\beta}_0) + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\lambda}^T} \hat{\boldsymbol{\lambda}} \\ &\quad + \frac{\partial Q_{2n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) + O_p(n^{-1}) \\ \mathbf{0} &= Q_{3n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0) + \frac{\partial Q_{3n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} (\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) + O_p(n^{-1}). \end{aligned}$$

From the last equation,

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0 &= - \left\{ \frac{\partial Q_{3n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \right\}^{-1} Q_{3n}(\boldsymbol{\beta}_0, \mathbf{0}, \boldsymbol{\theta}_0) + o_p(n^{-1/2}) \\ &= \mathbf{H}^{-1} Q_{3n} + o_p(n^{-1/2}), \end{aligned}$$

thus

$$\begin{aligned} \mathbf{0} = & \begin{pmatrix} Q_{1n}(\beta_0, \mathbf{0}, \theta_0) \\ Q_{2n}(\beta_0, \mathbf{0}, \theta_0) \end{pmatrix} + \begin{pmatrix} \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} & \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} & \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} \end{pmatrix} \cdot \begin{pmatrix} \hat{\lambda} \\ \hat{\beta}_{\text{EL2}} - \beta_0 \end{pmatrix} \\ & + \begin{pmatrix} \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \theta^T} (\mathbf{H}^{-1} Q_{3n}) \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \theta^T} (\mathbf{H}^{-1} Q_{3n}) \end{pmatrix} + o_p(n^{-1/2}). \end{aligned} \quad (\text{C.2})$$

Noting that

$$\begin{aligned} \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial g_i(\beta, \theta) / \partial \beta^T}{1 + \lambda^T g_i(\beta, \theta)} + o_p(n^{-1/2}) \xrightarrow{p} \begin{pmatrix} -\mathbf{S} \\ \mathbf{J}^T \end{pmatrix}, \\ \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} &= \frac{1}{n} \sum_{i=1}^n \frac{-g_i(\beta, \theta) g_i(\beta, \theta)^T}{(1 + \lambda^T g_i(\beta, \theta))^2} + o_p(n^{-1/2}) \xrightarrow{p} \begin{pmatrix} -\mathbf{S} & \mathbf{0} \\ \mathbf{0} & -\Omega \end{pmatrix}, \\ \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \theta^T} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial g_i(\beta, \theta) / \partial \theta^T}{1 + \lambda^T g_i(\beta, \theta)} + o_p(n^{-1/2}) \xrightarrow{p} \begin{pmatrix} \mathbf{0} \\ -\mathbf{H} \end{pmatrix}, \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} &= \frac{1}{n} \sum_{i=1}^n \frac{\lambda^T \partial^2 g_i(\beta, \theta) / \partial \beta \partial \beta^T}{1 + \lambda^T g_i(\beta, \theta)} + o_p(n^{-1/2}) \xrightarrow{p} \mathbf{0}, \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \theta^T} &= \frac{1}{n} \sum_{i=1}^n \frac{\lambda^T \partial^2 g_i(\beta, \theta) / \partial \beta \partial \theta^T}{1 + \lambda^T g_i(\beta, \theta)} + o_p(n^{-1/2}) \xrightarrow{p} \mathbf{0}, \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial g_i(\beta, \theta) / \partial \beta}{1 + \lambda^T g_i(\beta, \theta)} + o_p(n^{-1/2}) \xrightarrow{p} \begin{pmatrix} -\mathbf{S} & \mathbf{J} \end{pmatrix}, \end{aligned}$$

we have

$$\mathbf{0} = \begin{pmatrix} Q_{1n}(\beta_0, \mathbf{0}, \theta_0) + \begin{pmatrix} 0 \\ -Q_{3n}(\beta_0, \mathbf{0}, \theta_0) \end{pmatrix} \\ Q_{2n}(\beta_0, \mathbf{0}, \theta_0) \end{pmatrix} + \begin{pmatrix} \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} & \frac{\partial Q_{1n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} \\ \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \lambda^T} & \frac{\partial Q_{2n}(\beta_0, \mathbf{0}, \theta_0)}{\partial \beta^T} \end{pmatrix}$$

$$\cdot \begin{pmatrix} \hat{\lambda} \\ \hat{\beta}_{\text{EL2}} - \beta_0 \end{pmatrix} + o_p(n^{-1/2}) := \mathbf{v}_n + G_n \begin{pmatrix} \hat{\lambda} \\ \hat{\beta}_{\text{EL2}} - \beta_0 \end{pmatrix} + o_p(n^{-1/2}) \quad (\text{C.3})$$

with

$$\sqrt{n}v_n \xrightarrow{d} N(\mathbf{0}, \Sigma).$$

$$G_n \xrightarrow{p} G := \begin{pmatrix} -\mathbf{S} & \mathbf{0} & -\mathbf{S} \\ \mathbf{0} & -\Omega & \mathbf{J}^T \\ -\mathbf{S} & \mathbf{J} & \mathbf{0} \end{pmatrix}.$$

Thus

$$\sqrt{n} \begin{pmatrix} \hat{\lambda} \\ \hat{\beta}_{\text{EL2}} - \beta_0 \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, G^{-1}\Sigma G^{-1})$$

where the bottom-right block of $G^{-1}\Sigma G^{-1}$ is

$$\begin{aligned} \Sigma_2 = & (\mathbf{S} + \mathbf{J}\Omega^{-1}\mathbf{J}^T)^{-1} \{ \mathbf{S} - \mathbf{J}\Omega^{-1}\mathbf{U} - \mathbf{U}^T\Omega^{-1}\mathbf{J}^T + \mathbf{J}\Omega^{-1}(\Omega - \mathbf{V} - \mathbf{V}^T + \mathbf{W})\Omega^{-1}\mathbf{J}^T \} \\ & \cdot (\mathbf{S} + \mathbf{J}\Omega^{-1}\mathbf{J}^T)^{-1}. \end{aligned}$$

which is the same as $\text{ACov}\{\sqrt{n}(\hat{\beta}_{\text{EL0-2}} - \beta_0)\}$. We also know from Corollary 1 of Qin and Lawless (1994) that the efficiency of an estimator cannot decrease when adding an estimating equation, thus $\text{ACov}\{\sqrt{n}(\hat{\beta}_{\text{EL2}} - \beta_0)\}$ and $\text{ACov}\{\sqrt{n}(\hat{\beta}_{\text{EL0-2}} - \beta_0)\}$ must be less than or equal to the CML0 estimator which uses only one estimating function $\mathbf{s}_{c,\beta}(Y, \mathbf{X}, \mathbf{Z}; \beta)$.

We denote

$$\begin{aligned}
\mathbf{M} &:= -E\{R\partial \mathbf{s}_{c,\beta}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_0)/\partial \boldsymbol{\alpha}^T\}, \\
\mathbf{N} &:= -E\{R\partial \mathbf{u}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_0)/\partial \boldsymbol{\alpha}^T\}, \\
\mathbf{O} &:= -E\{\partial \mathbf{s}_\alpha(\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}^T\}, \\
\mathbf{Q} &:= E\{\mathbf{s}_{c,\beta}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\mathbf{s}_{c,\alpha}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^T\}, \\
\mathbf{R} &:= -E\{\partial \mathbf{s}_{c,\alpha}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\mathbf{h}(\boldsymbol{\theta}_0)\}
\end{aligned}$$

Then for EL3, the asymptotic variance by (1.12) is

$$\boldsymbol{\Sigma}_3 = \left[\begin{pmatrix} -\mathbf{S} & \mathbf{J}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} & \mathbf{H} \\ -\mathbf{M} & -\mathbf{N} & -\mathbf{O} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{S} & \mathbf{0} & \mathbf{M}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} & \mathbf{N}^T & \mathbf{V}^T \\ \mathbf{M} & \mathbf{N} & \mathbf{O} & \mathbf{N}^T \\ \mathbf{0} & \mathbf{V} & \mathbf{N} & \mathbf{W} \end{pmatrix}^{-1} \begin{pmatrix} -\mathbf{S} & \mathbf{0} & -\mathbf{M}^T \\ \mathbf{J} & \mathbf{H} & -\mathbf{N}^T \\ \mathbf{0} & \mathbf{0} & -\mathbf{O} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} \end{pmatrix} \right]^{-1} \quad (\text{C.4})$$

For EL4, the asymptotic variance by (1.12) is

$$\boldsymbol{\Sigma}_4 = \left[\begin{pmatrix} -\mathbf{S} & \mathbf{J}^T & \mathbf{0} & -\mathbf{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} & \mathbf{0} & \mathbf{H} \\ -\mathbf{M} & -\mathbf{N} & -\mathbf{O} & -\mathbf{O} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{S} & \mathbf{0} & \mathbf{M}^T & \mathbf{M}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} & \mathbf{N}^T & \mathbf{N}^T & \mathbf{V}^T \\ \mathbf{M} & \mathbf{N} & \mathbf{Q} & \mathbf{Q} & \mathbf{N}^T \\ \mathbf{M} & \mathbf{N} & \mathbf{Q} & \mathbf{O} & \mathbf{R}^T \\ \mathbf{0} & \mathbf{V} & \mathbf{N} & \mathbf{R} & \mathbf{W} \end{pmatrix}^{-1} \begin{pmatrix} -\mathbf{S} & \mathbf{0} & \mathbf{M}^T \\ \mathbf{J} & \mathbf{H} & -\mathbf{N} \\ \mathbf{0} & \mathbf{0} & -\mathbf{O} \\ -\mathbf{N} & \mathbf{0} & -\mathbf{O} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} \end{pmatrix} \right]^{-1} \quad (\text{C.5})$$

For EL5, the asymptotic variance by (1.12) is

$$\Sigma_5 = \left[\begin{pmatrix} -S & J^T & N & 0 \\ 0 & H & 0 & H \\ -M & -N & 0 & 0 \end{pmatrix} \begin{pmatrix} S & 0 & 0 & 0 \\ 0 & \Omega & N & V^T \\ M & N & Q - O & N^T - R^T \\ 0 & V & N - R & W \end{pmatrix}^{-1} \begin{pmatrix} -S & 0 & -M^T \\ J & H & -N^T \\ N & 0 & 0 \\ 0 & H & 0 \end{pmatrix} \right]^{-1} \quad (\text{C.6})$$

C.4 Additional simulation study 3: a logistic regression model with surrogate covariate

In this section, we have similar data generation procedure as Study 1 but with a surrogate covariate. We first generate (W, Z) from a bivariate normal distribution, with both variables having zero mean and unit standard deviation. The correlation of them is varied to reflect different degrees of similarities between the surrogate covariate and the real influential covariate. X is a categorized version of W , coded as 0,1,2 respectively for W values in the first, second and third tertile on break points $(-0.44, 0.44)$. We generate Y as a realization of $i = 1, \dots, N = 2000$ Bernoulli trials each with probability

$$P(Y_i = 1|Z_i) = \text{expit}(\beta_c + \beta_Z Z_i).$$

We take $\beta_0 = (-4, 1)$ which results in about 6% cases have $Y = 1$, and represents a rare disease incidence.

The indicator R_i to indicate whether the i -th subject is included in phase 2 sample is

Table C.1: Relative efficiencies for logistic regression models with surrogate covariates

Method	Relative efficiencies with respect to ML											
	$\rho=0.5$		$\rho=0.7$		$\rho=0.9$		$\rho=0.5$		$\rho=0.7$		$\rho=0.9$	
	$\beta_c = 0$	$\beta_z = 1$	$\beta_c = 0$	$\beta_z = 1$	$\beta_c = 0$	$\beta_z = 1$	$\beta_c = 0$	$\beta_z = 0.5$	$\beta_c = 0$	$\beta_z = 0.5$	$\beta_c = 0$	$\beta_z = 0.5$
CML0	60	83	58	67	56	58	66	87	63	75	56	63
CML-est	99	83	98	67	94	58	97	87	92	75	82	63
SW-est	99	83	98	67	94	58	97	87	92	75	82	63
EL5-est	100	95	99	93	93	96	99	98	101	99	98	100
CML-sat	99	91	84	94	58	83	98	89	102	81	93	75
SW-sat	99	84	91	84	93	68	98	89	101	81	93	75
EL5-sat	97	92	96	91	95	86	99	98	104	99	97	101
ML	100	100	100	100	100	100	100	100	100	100	100	100

also Bernoulli following

$$P(R_i = 1|Y_i, X_i) = \pi_1(Y_i; \boldsymbol{\alpha}) = \text{expit}(\alpha_c + \alpha_Y Y_i).$$

The nuisance parameter $\boldsymbol{\alpha}_0 = (-3.4, 3)$ so that we sample around 5% of the total phase 1 sample into phase 2, and those has $Y = 1$ and $Y = 0$ are of approximately equal size in phase 2. We consider a correct π model π -est which is a logistic regression with covariate Y , and another model π -sat to include a richer stratification of the phase 1 sample. Specifically, we stratify the combinations of (Y_i, X_i) into 4 strata, $(Y_i = 1, X_i = 0)$, $(Y_i = 1, X_i > 0)$, $(Y_i = 0, X_i = 0)$ and $(Y_i = 0, X_i > 0)$. This corresponds to a logistic model with covariates $(Y, V, Y * V)$ where $V = I(X > 0)$. We denote the two π models as “est” and “sat” and respectively name the corresponding estimators using either model.

The results of 500 runs of simulations of each setting are shown in Table C.1. We see that for a surrogate covariate which has high correlation with the original covariate X , the EL7 estimator significantly improved the estimation efficiency. The improvement decreases as the correlation between the surrogate and original covariate decreases, which is sensible. For a surrogate with less than 0.5 correlation with the original covariate, the meaning of this surrogate variable may be little.

C.5 Additional simulation study 4: normal linear regressions with varying covariate correlation

In this study, we follow the same data generation procedure as Section 4.3.2 with positive selection probability, namely $\boldsymbol{\alpha} = (0.3, 0.05, 0.5)$, but vary the correlation coefficient between \tilde{X} and Z , and also the coefficients in the main model. Specifically, we choose $\rho = 0.3, 0.5, 0.7$. The results of 500 runs of simulations with $\boldsymbol{\beta} = (0, 0.5, 0.5)$ and $\boldsymbol{\beta} = (0, 1, 1)$ are shown in Table. C.2 and C.3, respectively.

For both $\boldsymbol{\beta} = (0, 1, 1)$ and $(0, 0.5, 0.5)$, we observe similar comparisons between the estimators. When the π -est model is employed, we see that the EL4, EL5 estimators have significant improvement upon the SW estimator, which is about 30% decrease in the empirical SE's. When the π -sat model is employed, the SW estimator is already very efficient comparing to ML, thus the improvement is not as obvious. We do observe that Recalling the estimating functions of SW and the EL estimators, the EL4 and EL5 estimator has augmented the SW estimating functions with auxiliary \mathbf{u} and \mathbf{h} , thus more efficient than SW. However, we do not have any theoretical comparison between EL3 and SW. It is also confirmed here, that SW does better in the estimation of σ than EL3, while EL3 does better in the estimation of β_x . .

Table C.2: Relative efficiencies for a linear regression model, with $\beta_{c0} = 0$, $\beta_{z0} = 1$, $\beta_{x0} = 1$, $\sigma_0 = 2$.

Method	Relative efficiencies with respect to ML											
	$\rho=0.3$				$\rho=0.5$				$\rho=0.7$			
	β_c	β_z	β_x	σ	β_c	β_z	β_x	σ	β_c	β_z	β_x	σ
CML0	78	95	68	74	72	100	67	67	79	103	70	70
CML-est	81	95	68	77	74	100	67	67	80	103	70	70
SW-est	83	97	69	86	75	100	67	67	82	104	71	71
EL3-est	103	95	96	77	97	100	94	94	97	104	90	90
EL4-est	104	96	98	88	99	101	97	97	98	103	92	92
EL5-est	105	96	97	88	99	101	96	96	97	103	91	91
CML-sat	93	94	83	78	86	100	82	82	90	102	84	84
SW-sat	99	96	91	89	93	100	94	94	96	103	91	91
EL3-sat	103	95	96	77	97	100	94	94	97	104	90	90
EL4-sat	104	96	96	92	98	99	96	96	98	104	92	92
EL5-sat	104	96	95	92	98	100	95	95	98	104	92	92
ML	100	100	100	100	100	100	100	100	100	100	100	100

Table C.3: Relative efficiencies for a linear regression model, with $\beta_{c0} = 0$, $\beta_{z0} = 0.5$, $\beta_{x0} = 0.5$, $\sigma_0 = 2$.

Method	Relative efficiencies with respect to ML											
	$\rho=0.3$				$\rho=0.5$				$\rho=0.7$			
	β_c	β_z	β_x	σ	β_c	β_z	β_x	σ	β_c	β_z	β_x	σ
CML0	68	92	71	60	67	100	67	67	81	102	84	58
CML-est	72	92	71	63	71	100	68	61	83	101	85	60
SW-est	74	93	71	77	73	102	68	74	85	102	85	72
EL3-est	99	92	106	67	99	101	96	64	100	101	100	62
EL4-est	100	92	107	78	100	103	98	77	100	101	100	75
EL5-est	100	93	107	78	100	102	98	77	100	100	101	75
CML-sat	87	91	93	63	88	100	88	62	92	100	96	62
SW-sat	93	93	101	77	93	102	91	77	96	100	98	76
EL3-sat	99	92	106	67	99	101	96	64	100	101	100	62
EL4-sat	100	92	107	81	100	102	98	81	100	99	100	79
EL5-sat	100	93	107	80	99	102	97	81	100	100	100	79
ML	100	100	100	100	100	100	100	100	100	100	100	100

Glossary

AIPW Augmented inverse probability weighting [7](#)

ATE Average treatment effect [97](#)

BSS Basic stratified sampling [58](#)

CC Complete case [2](#), [4](#)

CDM Covariate dependent missing [3](#)

CML Conditional maximum likelihood [7](#)

DR Doubly robust [8](#)

GMM Generalized method of moments [12](#)

IPW Inverse probability weighting [6](#)

MAR Missing at random [3](#)

MCAR Missing completely at random [3](#)

MI Multiple imputation [4](#)

ML Maximum likelihood [4](#)

MNAR Missing not at random [3](#)

MR Multiply robust [16](#)

ODS Outcome-dependent sample(s) [9](#)

VPS Variable probability sampling [48](#)